

**INSIGHTS INTO ISOGENIC CLONAL FISH LINE
DEVELOPMENT USING HIGH-THROUGHPUT
SEQUENCING TECHNOLOGIES**

A THESIS PRESENTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By

Münevver Oral

MSc



**UNIVERSITY OF
STIRLING**

INSTITUTE OF AQUACULTURE

October 2016

Declaration

I hereby declare that this thesis has been composed entirely by myself. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions via personal communications. It has neither been accepted, nor submitted for any other degree or qualification. Except where specifically acknowledged the work described in this thesis is the result of my own investigations.

Candidate

Name: Münevver Oral

Sign:

Date:

Principal Supervisor

Name: Dr. David J. Penman

Sign:

Date:

List of manuscripts in preparation

Oral, M., Colléter, J., Bekaert, M., Taggart, J.B., Palaiokostas, C., McAndrew, B.J., Vandeputte, M., Chatain, B., Kuhl, H., Reinhardt, R., Peruzzi, S. and Penman, D.J. Gene-centromere mapping in meiotic gynogenetic European seabass.

Oral, M., Colléter, J., Bekaert, M., Taggart, J.B., Bartie, K., Taggart, J.B., McAndrew, B.J., Vandeputte, M., François, A., Vergnet, A., Chatain, B. and Penman, D.J. Genome-wide verification of isogenicity of clone founders (G1) in European seabass (*Dicentrarchus Labrax*) through ddRADseq.

Oral, M., Taggart, J.B., Wehner, S., McAndrew, B., Penman, D., Fjellidal, P.G. and Hansen, T. Verification of isogenic nature of clonal lines in the Atlantic salmon (*Salmo salar*) through ddRADseq.

Presentations in conferences

- **September 2016** – EAS 2016, Europe Aquaculture Conference 2016, Edinburgh, Scotland, UK, oral presentation entitled as “*Genome-wide verification of isogenicity of clone founders (G1) in European seabass (Dicentrarchus labrax) through ddRADseq*”.
- **November 2015** – Institute of Aquaculture, Lunchtime Seminars, Stirling, UK, oral presentation entitled as “*Insights into isogenic clonal lines through the use of Next generation Sequencing (NGS) technologies*”.
- **June 2015** – The International Symposium on Genetics in Aquaculture XII, ISGA XII, Santiago de Compostela, Spain, oral presentation entitled as “*Verification of Isogenic Nature of Clonal Lines in the Atlantic salmon (Salmo salar) through ddRADseq*”.
- **February 2015** – Biennial Post Graduate PhD conference of University of Stirling, Stirling, UK, oral presentation entitled as “*Linkage mapping in meiotic gynogenetic European seabass (Dicentrarchus labrax) and the development of isogenic clonal lines in fish*”.
- **October 2014** – EAS 2014, Europe Aquaculture Conference 2014, San Sebastián, Donostia, Spain, oral presentation entitled as “*A SNP map of the European seabass genome based on meiotic gynogenetic family*”.
- **May 2013** – Post Graduate Research Conference of University of Stirling, Stirling, UK, poster presentation.

Acknowledgement

It is my pleasure to thank all those people who made this project possible.

I would like to express sincere gratitude to my principle supervisors Dr. David J. Penman and Professor Brendan J. McAndrew for their continuous guidelines, advices, support and most importantly encouragements throughout my PhD. Their office doors were always wide open to me for any questions and/or discussions. I particularly value their rapid and high quality feedback during the last quarter of the project. Their trust in me has helped shaping an independent researcher out of me. Throughout my PhD journey, I had privilege to work with Dr. John B. Taggart whose guidelines in lab were most valuable. He was always there, providing solutions and supporting every step of the way. Similarly, my sincere thanks to Dr. Michaël Bekaert who has introduced me into fascinating world of bioinformatics and data analysis. I will definitely miss the most insightful and fun conversations with both of you. I am also grateful to technicians in molecular lab; Jacque, Kerry, Sarah-Louise for the help in lab work for any moment needed. A warm thank you also goes to the all staff in the Institute of Aquaculture, I enjoyed chatting every one of you during coffee and lunch breaks in K1, especially Keith Ranson for his help in tropic aquarium. My thanks to all project partners Drs. Julie Colléter, Marc Vandeputte and Tom Hansen for excellent contribution during my PhD.

Particularly during the course of my PhD, the support that I received from friends has been really important to me. You all made me feel like home and made my Stirling experience more joyful and enriching. Greta, Beatrix, Joanna, Željka, Christoforos, Christos, Phelly, Antonis, Emily, Marie, Sean, Aisha, Jamilla, Taslima, Khalfan, Mikey, Gustavo, Andrew, Suleiman, Eric & Sara, Luisa, Ying, Udin, Winarti, Houra, and you Marion.! Special thanks to Christoforos for organising parties to keep the motivation up in long cold winter nights and Nicolas (my Latin teacher) & Stylianos..I am very happy to meet you guys. My heartfelt thanks goes for my beloved flatmates whom I call as *my international family*. Your company will be missed the most. "Greta, Christoforos, Christos and Benji" till we meet next time somewhere exotic with more sun shine.!! Ayşegül, Betül, Kenan, Tuğba varlığınız bana hep güç verdi.

The last but not the least, my greatest thanks goes to my family who has always supported and believed in me. Without knowing I had a great family to go back to – I would not have got this far! Thanks millions mum, dad and brother, Aydın.

Regarding funding, I am most grateful to Turkish government (1416/YLSY) Ministry of Education who have financed my PhD and made this project possible. This project also received money from European Union, as a workpackage of AquaExcel (FP7 and follow up Horizon²⁰²⁰).

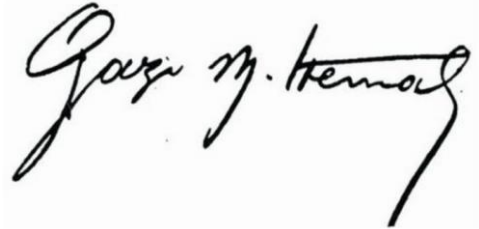
Zor diyorsun, zor olacak ki imtihân olsun..

Mevlânâ Celâleddîn-i Rûmî

Hayatta en hakiki yol gösterici ilimdir, fendir.

İlim ve fennin dışında yol gösterici aramak

gaflettir, dalalettir, cehalettir..

A handwritten signature in black ink, reading "Gazi M. Kemal". The signature is written in a cursive style with a long, sweeping tail on the letter 'l'.

AİLEM'e..

Abstract

Isogenic clonal fish lines are a powerful resource for aquaculture-related research. Fully inbred individuals, clone founders, can be produced either through mitotic gynogenesis or androgenesis and a further generation from those propagates fully inbred clonal lines. Despite rapid generation, as opposed to successive generation of sibling mating as in mice, the production of such lines may be hampered due to (i) potential residual contribution from irradiated gametes associated with poorly optimised protocols, (ii) reduced survival of clone founders and (iii) spontaneous arisal of meiotic gynogenetics with varying degree of heterozygosity, contaminating fully homozygous progenies.

This research set out to address challenges and gain insights into isogenic clonal fish lines development by using double-digest RADseq (ddRADseq) to generate large numbers of genetic markers covering the genome of interest.

Analysis of potential contribution from irradiated sperm indicated successful uniparental inheritance in meiotic and mitotic gynogenetics European seabass. Exclusive transmission of maternal alleles was detected in G1 progeny of Atlantic salmon (with a duplicated genome), while G2 progenies presented varying levels of sire contribution suggesting sub-optimal UV irradiation which was undetected previously with 27 microsatellite markers. Identification of telomeric markers in European seabass, with higher recombination frequencies for efficient differentiation of meiotic and mitotic gynogenetics was successful, and a genetic linkage map was generated from this data. One clear case of a spontaneous meiotic gynogenetic fish was detected among 18 putative DH fish in European seabass, despite earlier screening for isogenicity using 11 microsatellite markers. An unidentified larval DNA restriction digestion inhibition mechanism observed in Nile tilapia prevented the construction of SNP-based genetic linkage map.

In summary, this study provides strong evidence on efficacy of NGS technologies for the development and verification of isogenic clonal fish lines. Reliable establishment of isogenic clonal fish lines is critical for their utility as a research tool.

Table of Contents

Declaration	II
List of manuscripts in preparation	III
Acknowledgement	IV
Abstract.....	VI
Table of Contents	VII
List of Figures	XI
List of Tables.....	XIII
Abbreviations and Acronyms	XIV
Chapter 1	18
General introduction	18
1.1 State of aquaculture as a growing industry.....	19
1.2 Fish for scientific research.....	23
1.3 Uniparental reproduction and isogenic clonal lines	26
1.3.1 Chromosome set manipulations	28
1.3.1.1 Gynogenesis.....	30
1.3.1.2 Androgenesis	36
1.4 Isogenic clonal lines in fish research	39
1.4.1 Sex determination.....	42
1.4.2 Linkage and gene-centromere mapping.....	45
1.4.3 QTL mapping.....	47
1.4.4 Genetic and genomic resources.....	51
1.4.5 Limitations of isogenic clonal lines	53
1.5 Duplicated fish genomes and their complications	58
1.6 Next Generation Sequencing (NGS) technologies.....	61
1.6.1 Restriction based platforms (RADseq & ddRADseq).....	64
1.7 Aims and the objectives of the thesis	65
Chapter 2	68
General Material and Methods	68
2.1 General information	69
2.1 General maintenance of the Nile Tilapia stock in Tropical Aquarium	69
2.1.1 Fish stock origin and regulation.....	69
2.1.2 General maintenance of stock.....	70
2.2 Production of haploid and meiotic gynogenetic Nile tilapia.....	71
2.2.1 Collection of gametes.....	72
2.2.2 UV irradiation of sperm.....	72
2.2.3 Fertilisation.....	76
2.2.4 Application of heat shock to fertilised eggs for meiotic gynogenetic production.....	77
2.2.5 Incubation of eggs and sampling	78
2.3 DNA extraction protocols	81
2.3.1 REALpure DNA extraction protocol.....	82

2.3.2 SSTNE DNA extraction protocol.....	83
2.3.3 DNA quantification and standardisation.....	84
2.4 ddRAD library protocol	86
2.4.1 Restriction digestion.....	87
2.4.2 Ligation adaptors.....	88
2.4.3 First purification step.....	90
2.4.4 Size selection	91
2.4.5 Second purification step.....	93
2.4.6 Enrichment of library	94
2.4.7 Amplicon clean up (third purification).....	96
2.4.8 AMPure magnetic beads (final purification).....	96
2.5 Data Analysis.....	99
2.5.1 Quality Control of raw data.....	99
2.5.2 Stacks Pipeline for building loci.....	107
2.6 Microsatellites	110
2.6.1 General information.....	110
2.6.2 Fluorescent primer tailing.....	110
2.6.3 Polymerase Chain Reaction (PCR).....	111
2.6.4 Genotyping.....	111
Chapter 3.....	113
Multilocus analysis of a meiotic gynogenetic family of European seabass genome using ddRADseq.....	113
Abstract	114
3.1 Introduction	115
3.2 Materials and Methods	117
3.2.1 Production of mapping family - Meiogynogenetics	117
3.2.2 DNA preparation.....	118
3.2.3 ddRAD library preparation and sequencing.....	119
3.3 Data Analysis.....	120
3.3.1 Genotyping ddRAD alleles	120
3.3.2 Genetic linkage map construction	121
3.3.3 Visualising physical position of markers and microsats from previous studies	122
3.3.4 Marker-centromere mapping.....	122
3.3.5 Comparison of genomic assembly with linkage maps.....	123
3.4 Results	124
3.4.1 ddRAD sequencing	125
3.4.2 Investigation of potential sire contribution.....	125
3.4.3 Construction of female genetic linkage map.....	125
3.4.4 Physical position of markers in seabass genome	128
3.4.5 Marker-centromere mapping	130
3.5 Discussion.....	134
3.5.1 Conclusion.....	138
Chapter 4.....	140
Genome-wide verification of isogenicity of clone founders (G1) in European seabass (<i>Dicentrarchus labrax</i>) through ddRADseq.....	140
Abstract	141
4.1 Introduction	142
4.2 Materials and Methods	147
4.2.1 Production of clone founders through mitotic gynogenetics	147
4.2.1.1 Overview.....	147
4.2.1.2 UV irradiation of sperm, pressure shock and husbandry	148
4.2.2 ddRAD library preparation and sequencing by synthesis.....	150

4.3 Data Analysis	154
4.3.1 Sequence Quality Control (QC)	154
4.3.2 SNP calling	154
4.3.3 Investigation of putative sire contributor loci	155
4.4 Results.....	156
4.4.1 ddRAD sequencing.....	156
4.4.2 Distribution of ddRAD alleles	159
4.4.3 Investigation of putative sire contributor loci	159
4.4.3.1 Meiotic gynogenetic detected in F6 family	163
4.5 Discussion	165
4.5.1 Conclusions.....	171
Chapter 5	173
Verification of isogenic nature of clonal lines in the Atlantic salmon (<i>Salmo salar</i>) through ddRADseq	173
Abstract.....	174
5.1 Introduction.....	176
5.2 Materials and methods	184
5.2.1 Production of Isogenic clonal fish lines	184
5.2.1.1 Overview	184
5.2.1.2 UV irradiation of sperm and pressure shock.....	185
5.2.2 ddRAD library preparation and sequencing	186
5.2.3 Microsatellites.....	188
5.3 Data Analysis	192
5.3.1 Sequence Quality Control (QC)	192
5.3.2 Genotyping ddRAD alleles	192
5.3.3 SNP calling in G1 and G2 families	193
5.3.4 SNP calling in haploid family	193
5.3.5 Distribution of polymorphic ddRAD loci.....	194
5.3.6 Initial investigation of putative sire contributor loci	194
5.3.7 Identification of multi-copy loci	195
5.3.8 Investigation of sire contribution with one-copy loci.....	195
5.3.9 Finding the position of PCR primers using NCBI-BLAST	196
5.4 Results.....	199
5.4.1 ddRAD sequencing.....	199
5.4.2 Distribution of the ddRAD alleles	199
5.4.3 Initial investigation of the putative sire contribution in duplicated genome of salmon.....	200
5.4.4 Removal of multi-copy loci	203
5.4.5 Comparison of three available genome assemblies (<i>Ssal_v1</i> , <i>Ssal_v2</i> and <i>Ssal_v4</i>) in salmon.....	204
5.4.6 Polymorphic one-copy ddRAD loci	206
5.4.7 Microsatellites.....	211
5.5 Discussion	216
5.5.1 Conclusion	223
Chapter 6	225
Restriction digestion inhibition observed in the early developmental stages of Nile tilapia (<i>Oreochromis niloticus</i>).....	225
Abstract.....	226
6.1 Introduction.....	226
6.2 Materials and Methods	228
6.2.1 Ethics statement	228
6.2.2 Experimental design to sampling.....	228
6.2.3 DNA extraction and quantification	231

6.2.4 ddRAD library preparation and sequencing.....	231
6.3 Results	233
6.3.1 The first ddRAD library	233
6.3.2 The second ddRAD library.....	234
6.3.3 The third ddRAD library constructed as a control from fin clips	237
6.3.4 Troubleshooting	238
6.3.4.1 Adaptor & barcode test	239
6.3.4.2 Purification kit test.....	239
6.3.4.3 Restriction enzymes specificity.....	240
6.3.4.4 Time series sampling	242
6.4 Discussion.....	246
6.4.1 Conclusion.....	249
Chapter 7	251
General Discussion & Future Research Directions	251
General Discussion	252
7.1 European seabass (<i>Dicentrarchus labrax</i>)	252
7.2 Atlantic salmon (<i>Salmo salar</i>).....	258
7.3 Nile tilapia (<i>Oreochromis niloticus</i>).....	262
7.4 The role of HTS technologies on the future of Isogenic clonal fish line development	263
7.5 The role of HTS technologies on the future developments of aquaculture genomics	264
7.5 General summary	267
References	270
Appendix	291

List of Figures

Figure 1.1: Schematic diagram of chromosome manipulation techniques via gynogenesis to produce haploids, meiotic and mitotic gynogenetics.....	31
Figure 1.2: The summary diagram of production of meiotic and mitotic gynogenetics with a summary table explaining the specific requirements for each procedure and the details of the resultant progeny.....	33
Figure 1.3: A schematic representation of the production of isogenic clonal fish lines in two subsequent generations in female homogametic species (XX/XY) through mitotic gynogenesis.....	41
Figure 1.4: Whole-genome duplication (WGD) events during eukaryotic evolution. 1R, 2R, and 3R indicate first, second, and third-rounds of WGD in vertebrate evolution.....	59
Figure 2.1: Counting the concentration of sperm using haemocytometer.....	74
Figure 2.2: UV cabinet unit used for the sperm genome irradiation.....	75
Figure 2.3: A schematic diagram for the production of control groups and meiotic gynogenetic groups.....	76
Figure 2.4: Egg incubation system set up for Nile tilapia in Tropic Aquarium at University of Stirling.....	79
Figure 2.5: Workflow comparison of the two ddRAD library protocols; the original protocol by Peterson et al. (2012) is on the left and, the house-modified protocol is on the right.....	85
Figure 2.6: Design of the adaptors used for the construction of ddRAD library and the structure of the ddRAD library fragment formed by initial ligation of a SbfI P1 adapter (5bp barcode is blue colour coded) and a SphI P2 adapter (5bp barcode is blue colour coded).....	88
Figure 2.7: A schematic diagram of size selection on the agarose gel.....	92
Figure 2.8: A diagram represents the initial process of getting library with the size of interest from the agarose gel for the later purification process.....	93
Figure 2.9: Test PCR to optimise enrichment of the library.....	94
Figure 2.10: The change in the purification efficiency of paramagnetic beads as the ratio of bead:DNA decreases.....	96
Figure 2.11: AMPure paramagnetic beads clean-up procedure.....	98
Figure 2.12: A schematic diagram to show expected fluctuations at the beginning of the sequence reads.....	102
Figure 2.13: A schematic diagram explaining the entire Stacks pipeline in two consecutive stages: <code>denovo_map.pl</code> and <code>ref_map.pl</code> pipelines for loci building and SNP calling starting from cleaning and de-multiplexing through <code>process_radtags</code> module.....	108
Figure 3.1: Sequencing and ddRAD-tag summary.....	123
Figure 3.2: Genetic linkage map of meiotic gynogenetic <i>D. labrax</i>	126
Figure 3.3: Physical map position of SNP markers that have been identified in the present study from meiotic gynogenetic <i>D. labrax</i>	128

Figure 3.4: Detailed example of mapping in a single sea bass linkage group (LG 11), illustrating the computed recombination fraction for 79 progeny.....	129
Figure 3.5: Example of mono-arm chromosome, LG 11. A) Correlation of physical location (Mbp) with the percentage heterozygosity and genetic linkage map distance (cM) with genome assembly (Mbp). B) Frequency distribution of markers based on percentage of heterozygosity.....	130
Figure 3.6: Example of bi-arm chromosome, LG 14. A) Correlation of physical location (Mbp) with the percentage heterozygosity and genetic linkage map distance (cM) with genome assembly (Mbp). B) Frequency distribution of markers based on percentage of heterozygosity.....	131
Figure 3.7: Example of ambiguous chromosome, LG 7. A) Correlation of physical location (Mbp) with the percentage heterozygosity and genetic linkage map distance (cM) with genome assembly (Mbp). B) Frequency distribution of markers based on percentage of heterozygosity.....	131
Figure 3.8: Frequency distribution of marker-centromere distances.....	132
Figure 4.1: The pedigree of the samples used in the study.....	152
Figure 4.2: Sequencing and ddRAD-tag summary. Details of the number of reads before and after filters (orange disk) followed by the reconstructed number of ddRAD loci after filtering. The final number represents the polymorphic loci (markers) available per family after removing missing genotypes.....	156
Figure 5.1: The schematic diagram of experimental design used in the present study.....	190
Figure 5.2: Sequencing and ddRAD-tag summary.....	197
Figure 5.3: Physical position of microsatellites markers used in the present study alongside with Norwegian microsatellites. (continued in the next Figure).....	213
Figure 5.4: Physical position of microsatellites markers used in the present study alongside with Norwegian microsatellites.....	214
Figure 6.1: Gel images of the first ddRAD library constructed in Nile tilapia.....	233
Figure 6.2: Gel images of the second ddRAD library in Nile tilapia.....	234
Figure 6.3: Diagram shows the results of one round of sequencing run on MiSeq in terms of reads that were produced.....	235
Figure 6.4: Gel images of the third (control) ddRAD library that was constructed from four fin samples in Nile tilapia.....	237
Figure 6.5: Troubleshooting steps carried out to identify the reason for reduced yield observed in ddRAD libraries constructed in Nile tilapia.....	238
Figure 6.6: Restriction digestion profile of DNA from 2 fins and 2 larvae.....	240
Figure 6.7: Restriction digestion profile of 2 fins and 2 larvae.....	241
Figure 6.8: Image adapted from Fujimura & Okada (2007) to demonstrate time series sampling stages carried out in Nile tilapia.....	243
Figure 6.9: Double-digest restriction digestion profile of time series sampling carried out in Nile tilapia on 1.1% agarose gel.....	244

List of Tables

Table 1.1: Summary of isogenic clonal fish lines produced.....	39
Table 2.1: Phred score quality score interpretation.....	100
Table 3.1: Meio gynogenetic <i>D. labrax</i> genetic linkage map.....	125
Table 4.1: Summary of overall alignment rate of samples against to reference genome assembly of <i>D. labrax</i> (dicLab_v1) and overall depth of coverage achieved per sample at the end of one sequencing run.....	157
Table 4.2: The distribution of ddRAD alleles in F1 and F3 families.....	160
Table 4.3: The distribution of ddRAD alleles in F4, F6 families and two orphans (MO-926 & 927).....	160
Table 4.4: Summary of putative sire contributor loci in F1 and F3 families.....	161
Table 4.5: Summary of putative sire contributor loci in F4, F6 families and two orphans (MO-926 & 927).....	161
Table 4.6: The distribution of female heterogametic markers in the F6 family.....	163
Table 4.7: Summary table of the number of putative mitotic gynogenetics produced and genotyped using a panel of 12 microsatellite markers initially, followed by screening using ddRADseq of individuals homozygous for the microsatellite panel in European seabass.....	164
Table 5.1: The pedigree of the samples.....	189
Table 5.2: Distribution of ddRAD alleles in G1 and G2 families.....	201
Table 5.3: Distribution of ddRAD alleles in haploid family.....	201
Table 5.4: Summary table of BLAST analysis among three versions of genome assemblies.....	202
Table 5.5: Detailed BLAST analysis output of Ssal_v1, Ssal_v2 and Ssal_v4 genome assemblies (NCBI) in G1 and subsequent G2 families.....	204
Table 5.6: Detailed BLAST analysis output of Ssal_v1, Ssal_v2 and Ssal_v4 genome assemblies (NCBI) in haploid family.....	204
Table 5.7: Detailed representation of all non-duplicated loci in G1 family.....	207
Table 5.8: The summary table of informative markers in G1 family (non-duplicated loci)	208
Table 5.9: The summary table of informative markers in G2 families (non-duplicated loci).....	209
Table 5.10: The summary table of informative markers in haploid family (non-duplicated loci).....	209
Table 5.11: Inheritance of microsatellite alleles (in bp) from outbred founders to G1 and G2 families.....	212
Table 6.1: Schematic diagram of the experimental design.....	228
Table 6.2: List of the samples used.....	229
Table 6.3: Schematic diagram of time series sampling regime carried out in a bi-parental Nile tilapia family.....	242

Abbreviations and Acronyms

Below is a list of the most commonly used abbreviations in the text. Other abbreviated terms are explained in the text.

Abbreviation	Meaning
$\mu\text{W}/\text{cm}^{-2}$	Milliwatts per square centimeter
‰	Per-mille, parts per thousands
3Rs	Reduce, Replace, Refine framework
ASCII	American Standard Coding for Quality Scores
ASPA	Animal Scientific Procedures Act
Benzocaine	ethyl-4- aminobenzoate, a kind of anaesthetic
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BR	Broad Range
ca. & c.	"approximately" (Latin)
cM	CentiMorgan
daf	Days After Fertilisation
ddRAD	Double-Digest Restriction Associated DNA
denovo_map.pl	No reference genome based perl scrips
DH	Doubled-Haploid
<i>dicLab</i>	<i>Dicentrarchus labrax</i> (European Seabass) genome assembly
DNA	Deoxyribonucleic acid
dsDNA	Double Stranded Deoxyribonucleic acid
ENA	European Nucleotide Archive
erg/mm^2	Erg per square millimeter
et al.	"and others" (Latin)
EtBr	Ethidium bromide
EtOH	Ethanol
e-value	Expectation value
FAO	Food and Agriculture Organization (United Nations)
FCR	Feed Conversion Rate
FDA	Food and Drug Administration of US government

G0	Outbred parental fish
G1 fish	First Generation Mitotic Gynogenetic (clone founders)
G2 fish	Second Generation Gynogenetic, Putative Isogenic Clonal Fish
haf	Hours after fertilisation
HMM	Hidden Markow Model
HMW	High Molecular Weight
HO	Home Office, UK
HS	High Sensitivity
HTS	High Throughput Sequencing Technologies
HW	Hardy Weinberg equilibrium
ID	Identity
IMR	Institute of Marine Research, Norway
INRA	National Institute of Agricultural Research, France
IoA	Institute of Aquaculture, University of Stirling, UK
LG	Linkage Group
Lh ⁻¹	Liters per hour
LHRHa	Luteinizing Hormone-Releasing Hormone analog
LOD	Logarithm of odds (base-10) value
MAS	Marker Assisted Selection
MFR	Modified Fish Ringer's Solution
mJ/cm ²	millijoule per square centimeter
Mmix	Master-Mix
MHC	Major Histocompatibility Complex
MSV	Multisite Sequence Variant
mya	million years ago
NaAc	Sodium Acetate, also abbreviated CH ₃ COONa or NaOAc
NTC	No template control
N50	Statistical measure where shortest sequence lengths at 50% of the genome
NCBI	National Center for Biotechnology Information, Gene Bank
NGS	Next Generation Sequencing
nm	Nanometer
PCR	Polymerase Chain Reaction
PE	Paired-end
PIC	Polymorphic Information Content
PE	Polyethylene
PhiX DNA	PhiX virus control library
PIL	Procedure Personal Licence
PIT tag	Passive Integrated Transponder

PPCL	Potential Paternal Contributor Loci
PPL	Procedure Project Licence
PSC	Potential Sire Contribution
psi	per square inch
PSSD	Probable Small-Scale Duplication
PSV	Paralogous Sequence Variant
QC	Quality Control
qPCR	quantitative PCR
QTL	Quantitative Trait Locus
RAD	Restriction Associated DNA
RE	Restriction digestion
ref_map.pl	Reference genome based perl scrips
RNA	Ribonucleic acid
RT	Room temperature
SDS	Sodium Dodecyl Sulphate
seq	Sequencing
SGSS	Seabass Gamete Short term Storage
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
<i>Ssal</i>	<i>Salmo salar</i> (Atlantic salmon) genome assembly
SSTNE	Spermidine-Spermine-Tris-NaCl-EGTA
TA	Tropical Aquarium
TSA	Transcriptome Assembly Sequence
U	Enzyme Unit
UK	United Kingdom
UV	Ultraviolet
v/v	volume/volume
V _A	Additive genetic variance
V _D	Dominance-related genetic variance
V _E	Environmental variance
V _G	Genetic variance
V _{GE}	Genetic-Environmental correlated genetic variance
V _I	Interaction-related genetic variance
W	Watt
WGD	Whole Genome Duplication
WZ/ZZ	Female heterogametic
xG	Relative Centrifugal Force (rcf)

XX/XY
y

Male heterogametic
The frequency of second meiotic division segregation

Chapter 1

General introduction

1.1 State of aquaculture as a growing industry

Aquaculture is the fastest growing food sector globally. As a result the world fish consumption *per capita* increased significantly from 9.9 kg in the 1960s to 14.4 kg in 1990s and almost 20 kg in 2013 with an annual growth rate of 6.9% (FAO, 2016). Global production of aquaculture amounted to 131 million tonnes, constituting 73.8 million tonnes of fish (49.8 million tonnes of finfish) and 16.1 million tonnes of molluscs, 6.9 million tonnes of crustaceans, 7.3 million tonnes of other aquatic animals and 27.3 million tonnes aquatic plants in 2014 (FAO, 2016). Thus farmed fish constitutes the three quarter of the total aquaculture production in volume. According to FAO's estimates, over 50 million tonnes more seafood is required so as to meet increasing demand of human consumption and growing population by 2050. However this is not an easy target to reach given the reality of climate change and increasing competition for natural resources. To meet such targets, the aquaculture industry has to increase production in sustainable ways. To this end, applying selective breeding programs, including integration of modern genetics technologies, can significantly accelerate the production through increased use of genetically improved stock in the aquaculture sector.

Although aquaculture has utilised traditional methods of selection (i.e: using bigger fish as broodstock) and/or domestication in carp species in Asia for decades, the establishment of advanced and sophisticated selective breeding programmes were started in species of prime commercial interest such as salmonids and tilapia (see reviews of Hulata, 2001 and Gjedrem, 2005) in 1970s with more advanced selections are being implemented at 1990s. The benefits of selective breeding have been recognised over the last decade where initial phase genetic breeding programmes generated high returns (Gjerdem et al., 2012; Gjerdem & Robinson, 2014). Thus, as of today many national and international level selective breeding programmes have been set up in aquatic species such as Channel

catfish *Ictalurus punctatus* (Dunham & Brummet, 1999), multi-species tilapia hybrid in Israel (Hulata et al., 1999), turbot *Scophthalmus maximus*, European seabass *Dicentrarchus labrax*, gilthead seabream *Sparus aurata* and others (see recent survey by Chavanne et al., 2016) so as to utilise modern and traditional breeding methods. The international funding available for such improvements (e.g. Fishboost project, EU project aim to improve productivity traits in six species of prime commercial interest in Europe) should help this goal to be achieved more rapidly. As of today only 8.2% of aquaculture production is based on genetically improved stocks globally (Neira, 2010; Rye et al., 2010) however this is expected to be higher, given this statement dates back to six years ago. Moreover, this proportion is highly diverse among fish species; e.g. commercial aquaculture production is heavily reliant on genetically improved stocks in Atlantic salmon, while in some other species of commercial interest such as Barramundi, Asian seabass, with a total global production of 71,581 tonnes in 2014 (FAO 2016), selective breeding has not been implemented, yet. Gjedrem, Robinson & Rye (2012) have estimated a great potential for almost a twenty-fold increase in aquaculture production globally, with one of the main limitations being availability of feed resources for fish species. The same authors suggested world aquaculture production could be doubled in thirteen years with an overall potential genetic gain of 12.5% per generation if selective breeding was applied in all farmed aquatic species.

Over the last two decades or more, modern genetic technologies such as chromosome set manipulations (e.g. triploid rainbow trout and Pacific Oyster production), cryopreservation of the gametes, sex reversal to produce mono-sex fast growing commercial populations (e.g. mono-sex rainbow trout and Atlantic halibut commercial production), genome and quantitative trait loci (QTL) mapping (e.g. commercial Marker-Assisted Selection for resistance to infectious pancreatic necrosis in Atlantic salmon)

(Houston et al., 2008, 2012; Moen et al., 2009) have been transformed from experimental stages to being a part of regular commercial aquaculture production. Such technologies are advancing the aquaculture industry mainly in developed countries; yet more work needs to be done in developing countries where most of the seafood production comes from.

A promising breakthrough for aquaculture yet to be implemented is a technology called genomic selection. This technology offers higher accuracy for an individual even prior to phenotyping and improves selection responses significantly. This mode of selection is based on first estimating the effects of dense genetic markers in a test population and then uses such information to predict breeding values of selection candidates (Nirea et al., 2012). In the aquaculture sector, where breeding programs are still in their infancy for some species (such as mass spawners, Asian seabass) and more sophisticated for others (e.g: Atlantic salmon), genomic selection clearly provides an avenue for increased genetic gain and a direction to go after what has been achieved in terrestrial animals (Liu, 2011). However, as the underlying genomic technology is heavily based on simulations, predictions works better under certain limited assumptions such as equally spaced QTL centered between two markers. Therefore this technology requires implementation of the refined methodologies in production systems and validation of cost-effectivity before maximising the benefits of genomic selection in fish species.

Modern DNA marker technologies allow efficient genetic linkage mapping. Knowledge gained on linkage between markers is applied identifying genomic regions that are associated with QTL. Tightly linked markers can then be utilised in MAS. QTLs are involved with the expression of a gene of interest for a trait (e.g. growth, fillet quality, reproductive traits, specific disease resistance) thus defining markers closer to QTL allows fast selection of candidates carrying with superior genotypes for the selected trait.

A breakthrough application of a QTL associated with a disease resistance was recently identified in Atlantic salmon by two independent research groups in Scotland and Norway (Houston et al., 2008, 2012; Moen et al., 2009). A major gene was identified and the knowledge gained from these researches was applied to the industry (the first generation of QTL-innOva® IPN eggs commercialised by AquaGen proved their defence mechanism toward the virus throughout their life cycle and performed well under commercial conditions).

Production of mono-sex male tilapia is desirable as a means of controlling reproduction in Nile tilapia (e.g. in pond culture) where females can reproduce every few weeks, thus creating overcrowding. Therefore tilapia farmers desire to produce all male mono-sex populations so as to delay or prevent sexual maturation. As YY tilapia (“supermales”) can be produced, this offers a tool for establishing a broodstock for production of mono-sex male commercial populations. Fertilising gametes of normal XX female should result in 100% XY mono-sex male production, which can be used for commercial population in Nile tilapia (Hulata, 2001). However there are problems in obtaining close to 100% male with this technique, due to complexities in sex determination. Given the widely-used hormonal sex reversal in Nile tilapia so as to control reproduction, this technique is used in only limited capacity. However the potential for commercial use may be increase in the future with the pressure of reducing use of steroid hormones in the food chain. Production of YY tilapia has been practised in commercial scale in Asia such as Philippines (since 1995), Thailand (1997), China and Vietnam (records of personal communication with Mair, 2000).

Despite ongoing effort into production of genetically modified transgenic fish, this has hardly made an impact in the aquaculture industry, mainly due to public concerns and long approval procedures. However, recently genetically engineered transgenic salmon,

known as AquAdvantage®, has been approved for human consumption by US Food and Drug Administration (FDA) on November 2015. Although discussions are still going on regarding human consumption concerns, labelling and allergens (Smith et al., 2010), AquAdvantage® constitutes the first genetically engineered animal food approved for human consumption. Although yet to make a significant impact on aquaculture industry, these fish show two-fold faster growth, sterility (due to induced triploidy: no concerns of wild population mating) and reduced FCR.

1.2 Fish for scientific research

Fish constitute one of the largest and most diverse groups in vertebrates: over half of all vertebrates are fishes (Froese & Pauly 2014). Great diversity exists between species, which combined with their varied habitats makes them tremendously attractive tools for studying a wide range of disciplines including ecology, developmental biology, behaviour, nutrition, physiology, anatomy, genetics and evolution.

The use of fish in scientific research is increasing globally. This is mainly due to rapid expansion of the aquaculture sector as well as increasing perception of using fish as a model to mammalian research in both fundamental research and drug testing. It is thus possible to decipher gene regulations or gain insights into human health and disease by exploiting fish genomes (Ahituv et al 2004). For example a recent study revealed a high homology (70%) between the zebrafish and human genomes, with remarkable similarity by sharing almost the same genes (Howe et al 2015). However understanding of their biology cannot be accomplished in the absence of experimentation with live fish.

Most fish species have high fecundity meaning that large numbers of gametes can be collected in a single spawning event. A large diversity is observed in number of gametes among fish species. For example an average fecundity for salmonids is around 10^3

eggs/kg while the number can go up to $>10^5$ eggs/kg in cyprinids. Plasticity of phenotypes is a well-known phenomenon in fish species in which the same genotype can possess many possible phenotypes depending on mainly environmental factors. For example, phenotypic sex can be changed by administering a temperature regime or hormonal treatment (depending on species) during the small window of the labile period.

A prerequisite for high quality experimental designs are:

- (i) To achieve high replicability with low variation between replicates
- (ii) To attain high reproducibility with robust results from different laboratories
- (iii) To accomplish high repeatability due to low variation between assays performed within the same laboratory (Dave 1993).

As of today most fish experiments still rely heavily on using outbred stocks with an exception of zebrafish where inbred lines are highly facilitated for experimental use. This not only decreases reproducibility in the experiments performed but also requires more animals to be used to achieve statistically powerful results.

The prerequisite for a good model organism to test any factor (e.g: response to a newly developed drug or chemical) is the sensitivity that the model organism reflects in response to the substance, even to minor dose changes. In this regard isogenic lines offer the best experimental tool due to minimal variation observed within the same isogenic animal line and maximum variation between different isogenic lines. Uncontrolled variation due to genetic and/or environmental factors or sex differences however reduces treatment effect and power to detect the effect of specific treatments. Experiments involving utilising isogenic lines increases the power since such animals will respond better to any changes in the experimental condition undertaken due to increased variation within the lines (Festing, 1995; Festing & Altman, 2002). To this end utilising isogenic animals derived from multiple strains decreases the use of many outbred derived from different

origins thus are of significant interest within the framework of the 3Rs (Replace, Reduce and Refine) concept by many national and international legislation bodies regulating the use of animals in scientific procedures (“NC3Rs”, 2016). For example in UK the use of animals in scientific experiments is regulated under the Animals (Scientific Procedures) Act 1986 (ASPA) which has recently been revised to be compatible with European Directive 2010/63/EU and its associated code of practise in animal care. This would further reduce the chance of animals being resistant to compound under investigation by utilising multiple lines, assuming the lines have not been set up by populations previously exposed to the compound. Therefore such experimental designs are more likely to measure more specific response of the genetic variation among different isogenic lines (Festing, 1992,1995,1999).

As Heston clearly stated on the discussion of utilising multi-strain isogenic experimental design over outbred stocks:

“Yet the question is sometimes asked, why not use genetically heterogeneous stock mice so the results will be more applicable to the genetically heterogeneous human population? The answer is that we are not trying to set up a model with mice exactly comparable with humans. This would be impossible because mice and men are different animals. What we are trying to do is to establish certain facts with experimental animals and this can be done, or done more easily, when the genetic factors are controlled. Once the facts are established we then, with much common sense, see how the facts can be related to man. When genetic variability is desired this can be obtained in the highest degree by using animals of a number of inbred strains. This variation between strains is usually much greater than is found in animals of a non-inbred stock which actually may be rather uniform although more variable than an inbred strain” (Heston 1968). This remains to be true with further supporting evidence after almost five decades.

Recently the need for genetically standardised fish lines has been reviewed by Grimholt et al. (2009). They stated that most results are not reproducible among different sites within the same species due to supply of large number of fish from various local breeders and national breeding companies. Therefore these authors highlighted research outcomes that could be facilitated by the establishment of isogenic clonal fish lines in Atlantic salmon as a species of prime importance for aquaculture in the north Atlantic.

1.3 Uniparental reproduction and isogenic clonal lines

As a non-Mendelian form of inheritance, uniparental reproduction refers to the transmission of the genome of only one parent to resultant progeny which would generate progeny containing all genes derived from either maternal or paternal source. Such reproduction techniques can be applied to produce isogenic clonal lines.

Isogenic clonal lines as the name suggests are populations of genetically identical and completely homozygous individuals. These can either be seen in nature, in rare cases, as in reptiles or amphibians or can be produced by applying experimental manipulations (Robinson & Thorgaard 2011). The term “clone” refers to no genetic differences among progeny and “line” corresponds to a set of genetically related individuals which are maintained under specific breed identification.

Isogenic clonal fish lines are remarkably similar to inbred lines of animals which have been extensively employed mostly in mice and plants, in the form of “recombinant inbred lines” (Beck et al., 2000; O’Neill et al., 2008). Inbred lines requires approximately 20 or more generations of full-sib matings (i.e. brother x sister) so that the inbreeding coefficient reaches up to almost 100% ($F=0.986$).

Although the production process takes long time, the utility of using inbred lines in research has been well documented (Beck et al., 2000). One of the most significant

outcomes using inbred lines was the discovery of H2, the major histocompatibility complex (MHC) of the mouse by Dr. George D. Snell, who later won Nobel Prize with the discovery. The significance of his work placed the foundation for dissecting the mechanism of transplantation in experimental animals and its ultimate transfer to humans. He developed the first ever congenic strains of mouse, which are in use even today. Furthermore he came up with the methodology of backcrossing, which is still an important tool in genetic mapping studies to dissect complexities of genomes of interest (Snell et al., 1976). Another significant example is the discovery of monoclonal antibodies by Milstein & Köhler in treatment of diverse diseases with a significant impact on medical research (Köhler & Milstein, 1975). None of these studies would be feasible without the availability of inbred strains of mice where increased homozygosity simplified the complex genetic analysis by eliminating the complexity of multiple alleles at a locus under investigation.

In fish, due to external fertilisation, gametes can be manipulated artificially. This enables researchers to manipulate the chromosome sets with high flexibility (Gjedrem, 2005) and propagate isogenic clonal lines in two subsequent generations (Dunham, 2004). Individual fish possessing inbreeding coefficient of $F=100\%$ can be produced in one generation either through mitotic gynogenesis or androgenesis using gametes from outbred fish. These are potential clone founders (isogenicity needs to be verified) and each progeny represents a unique genotype. Once the first generation fish reach maturity, a second round of gynogenesis or androgenesis is applied to the gametes of the resultant progeny of the first generation. This time however, outstandingly from the first generation, *clones* of the same genotypes descendent from each unique isogenic clone founder are produced ($F=100\%$). This rapid two-generation approach of producing isogenic clonal fish lines is strategically valuable in two ways: (i) by accelerating the

production of such lines in species with longer generation times, (ii) fast “fixation” of the desired genotype for a variety of interesting genotypes. Once isogenic clone founders (1st generation) and/or isogenic clonal fish lines (2nd generation) have been produced, a population of genetically identical but outbred fish can be produced by crossing between isogenic lines, termed outbred clones. Outbred clones possess a genetically identical genome, i.e. uniformly heterozygous. Bongers et al. (1997c) used outbred clones for the genetic analysis of testis development in common carp and verified suitability of such crosses for experimental animal models. If gametes from such outbred clones are used for another round of androgenesis or gynogenesis to produce new inbred forms involving the genomes of two clonal lines, they are called *recombined clonal lines* (Komen & Thorgaard, 2007). The terminology for such recombinants is depended on the organism that they are propagated. For example, the term recombinant inbred lines is used for plants while the term for recombinant inbred strains is used in mice.

1.3.1 Chromosome set manipulations

Chromosome set manipulations used in many aquatic species result in alterations of ploidy level such as haploids, triploids, tetraploids or diploids with uniparental inheritance (gynogenetics and androgenetics). These manipulations are not considered as genetic modifications (Migaud et al., 2013). According to EU regulations (Directive 90/220/CEE of April 23 of 1990), tetraploids or any other products of ploidy manipulations (i.e: triploids, gynogenetics or androgenetics) are not classified as Genetically Modified Organisms (GMOs).

These techniques have been primarily used for (i) production of polyploids, mostly triploids, to achieve sterility and continued growth with direct application to the aquaculture industry; and (ii) production of isogenic clonal fish lines for research

purposes (i.e: elucidating sex determination, genetic linkage mapping, QTL mapping). The use of chromosome set manipulations have been reviewed in detail in several studies (Thorgaard 1986; Dunham 2004; Overturf 2009) including the most recent one by Komen & Thorgaard (2007) on the development and the use of isogenic clonal fish lines for aquaculture related research.

Briefly these techniques involve two main steps: (i) the use of UV or gamma irradiation to inactivate the genetic material of maternal or paternal gametes prior to fertilisation; and/or (ii) suppression of meiosis II or mitosis I by using heat, pressure or chemical shocks. Since one of the parents will not contribute to the progeny genome following irradiation treatment, fertilisation will result in a haploid embryo with only one set of genetic material from the unirradiated parent. Following fertilisation a haploid embryo will be formed, however these are unable to survive long beyond the hatching stage. Therefore diploidy needs to be restored, achieved through a pressure or heat shock, to obtain viable progeny (Thorgaard 1983; Purdom 1993; Pandian & Koteeswaran 1998; Dunham 2004).

Recombination (exchange of genetic material between non-sister chromatids) occurs in the course of meiosis while the diploid set of chromosomes (one of which is of paternal origin and the other of maternal origin) duplicates then goes through two reductional divisions (during first meiosis four copies of chromosomes are reduced to two copies, followed by the final reduction of two sets to one in the second meiotic division) to produce gametes. During fertilisation, the merging of both gametes (one of which is paternal origin and the other is maternal origin) gives rise to the zygote and forms its natural diploid state. Fish eggs prior to fertilisation are still in the stage where two sets of maternal chromosomes are yet to undergo the second meiotic division, when the second polar body is excluded following fertilisation. In a typical bi-parental family, the second

polar body is excluded from the developing embryo following fertilisation by the sperm. Then single-celled zygote starts going through exponential cell divisions to form the multicellular embryo. The external fertilisation that fish present allows researchers to manipulate chromosome sets or alter ploidy levels in a relatively straightforward way (Penman & McAndrew, 2000; Lubzens et al., 2010).

1.3.1.1 Gynogenesis

Gynogenesis is a type of uniparental reproduction technique in which resultant progenies are produced possessing 100% maternal genome transmission. Therefore, sperm DNA has to be eliminated. This step is carried out either using X-ray, Ultra Violet light (UV) or gamma (γ) irradiation (Thorgaard 1983; Pandian & Koteeswaran 1998; Overturf 2009). Both X-rays and γ -irradiation have higher penetration level than UV does, thus they can effectively fragment the DNA into small sizes to avoid any parental contribution. However both these techniques require more expertise and investment than UV and very few laboratories have such facilities (Komen & Thorgaard, 2007). Unlike X-rays and gamma irradiation, UV irradiation attacks adjacent base pairs stimulating the induction of pyrimidine dimers (T-T, C-C, C-T in dsDNA; U-U, C-C, U-C in RNA), thus inhibiting the ability of the DNA polymerase enzyme to repair DNA damage in sperm (Durbeej & Eriksson 2002). UV irradiation is the methodology of choice in gametes of small size, particularly in sperm cells, due to its lower penetration power while ionizing radiation has generally been considered as a method of choice for irradiating larger volumes of sperm or large fish eggs, such as in salmonids, due to its high penetration power (Arai et al., 1979). Although the genomic material of sperm is fragmented the motility is not affected by the optimal irradiation treatment, thus such sperm are capable of initiating fertilisation and activating eggs to develop into embryos. This process leads to haploids in which the

maternal genome duplication needs to be induced to attain viable progeny. The only report of surviving a few gynogenetic haploids (until 50 days AF) was in *O. mossambicus* using UV irradiated sperm, however due to feeding difficulties haploid larva grew at one fourth of the rate of diploid counterparts (Varadaraj, 1993).

The diploid status of resultant progeny can be induced by applying a pressure or heat shock. Two different developmental stages can be targetted for production of diploid gynogenetic progeny; the naming comes after the interference applied, either during meiosis II or first mitosis (See Figures 1.1 and 1.2). The prevention of meiosis II is attained with the application of an *early* shock, shortly after fertilisation, to prevent exclusion of second polar body. This captures the results of every crossover during meiosis, thus varying levels of heterozygosity are produced in the resultant progeny depending on (i) the level of heterozygosity in the mother and (ii) the degree of recombination during meiosis. Mitotic gynogenetics, on the other hand, are produced with the suppression of the mitosis by administrating of a *late* shock, which causes an endomitosis. This leads to one haploid set of female chromosomes to be duplicated therefore 100% inbreeding is achieved in the resultant progeny. Such individuals are also referred as doubled haploids (DH). Each viable mitotic gynogenetic is a unique genotype derived from a singular dam haplotype.

Gynogenesis has been well studied in many finfish and shellfish species (Komen et al., 1991; Galbusera et al., 2000; Castro et al., 2003; Betotto et al., 2005; Tvedt et al., 2006; Nie et al., 2011) as reviewed by Komen and Thoorgaard, (2007). Gynogenesis has proven to be a very useful technique for the elucidation of sex determining systems, as the sex ratio of gynogenetic progeny would be informative for understanding sex determination under operation. Application of meiotic and mitotic gynogenesis, androgenesis as well as hormonal sex reversal and progeny testing was commonly applied in species where no

information was available regarding sex determination in the past. The outcomes of each procedure would add into the existing knowledge by increasing the understanding of sex determination where results from only one procedure could be misleading.

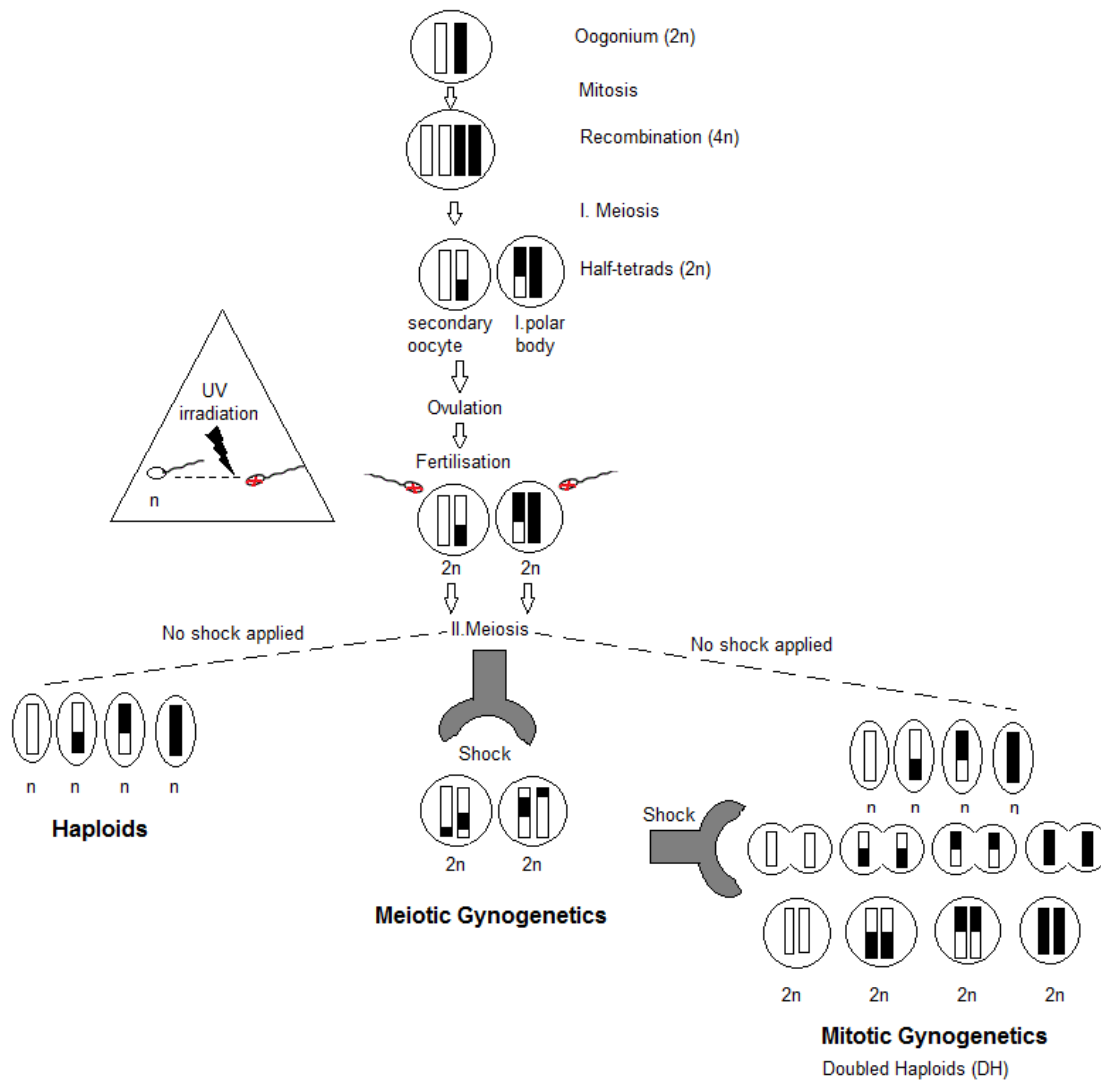
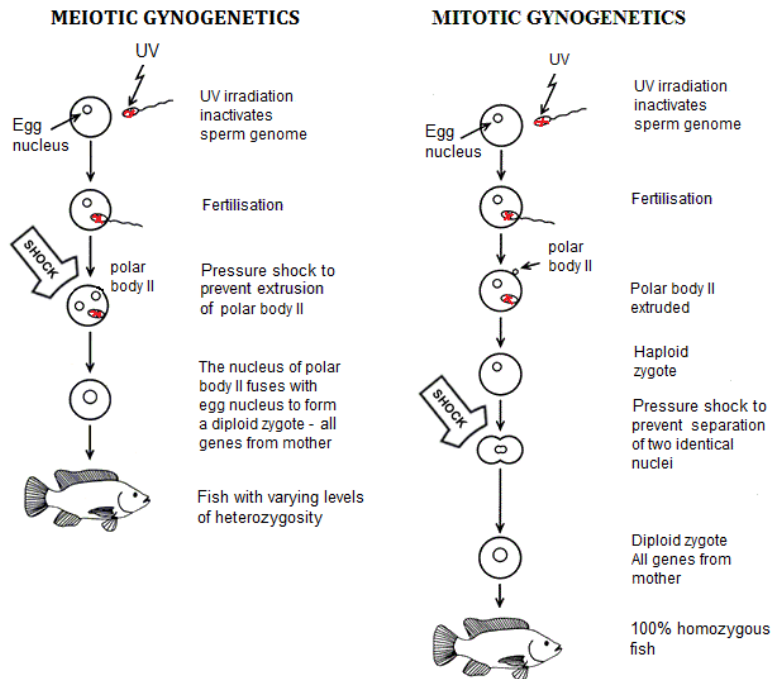


Figure 1.1: Schematic diagram of chromosome manipulation techniques via gynogenesis to produce haploids, meiotic and mitotic gynogenetics.

1.3.1.1.1 Mitotic Gynogenesis

The part of the procedure that sets mitotic gynogenesis apart from that used to induce meiotic gynogenetics is the shock, which targets the first mitosis of the developing embryo (Fig 1.1). This step is required to induce diploidy in the resultant offspring for the

production of viable progeny. A typical procedure of UV irradiation of sire sperm is followed by fertilisation and the exclusion of the second polar body, with the resultant haploid embryo carrying only one set of maternal chromosomes (Fig 1.2). This is when either a pressure or a heat shock is applied as the zygote undergoes the first cleavage, which leads to viable (diploid) mitotic gynogenetics. This is a difficult developmental phase to manipulate (see Fig 1.2), thus survival is significantly lower compared to meiotic gynogenetics. Survival comparison studies are particularly difficult to assess in gynogenetics as maternal effects and the physiological stage of eggs used can be variable among female breeders and in broader scale among different species. Quillet, (1994) studied survival, growth and reproductive traits in mitotic gynogenetics rainbow trout and reported 2% and 38% survival at two years old in mitotic gynogenetic group and control families produced from the same mother, respectively. The weight of DH mitotic gynogenetics was also detected 50% lower than controls. In another study, a total of 27.1% survival was reported at the hatching stage and 24% at yolk sac resumption in Nile tilapia mitotic gynogenetics compared to a mean survival of 79% at hatching and 70% in YSR in bi-parental control groups from the same female (Muller-Belecke & Horstgen-Schwark, 2000). Although experimental groups of mitotic gynogenetics are hard to produce, they are of considerable interest for establishing completely isogenic lines.



Groups	UV irradiation	Shock applied	Shock timing	Purpose	End result
Meiotic (2n) Gynogenetics	Yes	Early shock	Anaphase of Meiosis II	To prevent polar body II exclusion	Varying levels of heterozygosity: Loci closer to centromeric end will be homozygous while loci closer to telomeric end will be heterozygous
Mitotic (2n) Gynogenetics	Yes	Late shock	First Mitotic division	To suppress mitosis	100% homozygous and inbred

Figure 1.2: The summary diagram of production of meiotic and mitotic gynogenetics with a summary table explaining the specific requirements for each procedure and the details of the resultant progeny. The figure was adapted and modified from FAO website.

1.3.1.1.2 Meiotic Gynogenesis

The protocol for the production of meiotic gynogenetics is almost identical with that of induction of mitotic gynogenesis apart from the very important difference in the process where diploidy is restored by the application of an earlier shock (Fig 1.1). Such shock is applied to suppress expulsion of the second polar body following fertilisation of eggs with UV-irradiated sperm. This subsequently leads to a diploid zygotic genome which is of maternal origin only. However, prevention of second polar body exclusion from the

developing zygote captures the results of any crossover events in two of the four chromatids involved in homologous chromosome pairings during meiosis, which results in varying levels of heterozygosity in meiotic gynogenetic progeny (Fig 1.2). Given that crossover between a gene and the centromere is a relative measure of how far a gene and the centromere are located from one another, the frequency of heterozygotes in such progeny will be a direct reflection of recombination. For example in the absence of crossover, an informative locus segregating from a heterozygous female parent in the meiotic gynogenetic progeny will be homozygous for both alleles (e.g: *aa*, *bb*), while in the case of crossover between the centromere and the locus, the frequency of heterozygotes will provide a direct measurement of the distance between a gene under investigation and the centromere. For example if we assume we produced an experimentally propagated meiotic gynogenetic family with 100 offspring to locate the centromere on each linkage group, any locus closer to centromere will be represented with higher proportions of homozygotes in the progeny (e.g: 95 progeny out of 100 with *aa* and/or *bb* genotype or 100 progeny out of 100-if the locus is located almost on the centromeric region) while any loci distant from the centromere and adjacent to telomeric regions will be represented with mostly heterozygotes in the progeny (e.g. 87 loci out of 100 or 100 out of 100 with *ab* genotype). This analysis only takes into account the female heterogametic markers (with *ab* genotype) so as to observe the segregation of alleles; female homogametic markers are not informative. Overall, this artificially induced reproductive process enhances homozygosity, which will be a direct function of recombination taking place. However it is clear that such progeny will not be isogenic in all loci, but will be homozygous at any loci that are in centromeric regions (Devlin & Nagahama, 2002).

It is of prime interest to understand the major difference between both the production of meiotic and mitotic gynogenetics and the end results of each procedure for the understanding of current thesis throughout. Although meiotic gynogenetics are predominantly of interest for gene-centromere mapping, they can arise spontaneously in mitotic gynogenetic induction (see Chapter 4). As meiotic gynogenetics carry the results of crossover during the second meiosis they are partially heterozygous. The level of heterozygosity can be quite diverse in meiotic gynogenetics, e.g. $F = 55-79\%$ inbreeding coefficient was reported in carp species (Reddy 1999). Meiotic gynogenetics with varying level of heterozygosity need to be detected and eliminated from isogenic mitotic gynogenetic fish during the production of isogenic clonal fish lines.

1.3.1.2 Androgenesis

Androgenesis is type of uniparental reproduction technique in which resultant progeny possess 100% paternal nuclear genome transmission (with maternal origin mtDNA). It can be achieved artificially by inactivating the nuclear genomic content of the eggs with various methods such as ionising radiation or UV rays (see reviews: Pandian & Koteeswaran, 1998; Arai, 2001). Since no nuclear genomic contribution is received from the eggs, haploid embryos carrying only parental nuclear DNA are produced. Such haploids however are not viable and suffer from twisted body and curved tail alongside, thus they die around hatching stage. A pressure or heat shock is applied at the first cleavage (mitotic cell division) to restore diploidy and produce viable androgenetic progeny. Such shocks target first mitosis and since only one set of paternal chromosomes is duplicated the resultant progeny will be 100% inbred. This used to be a methodology of choice for species where there was no or little information available in species of interest. Thus outcomes of a range of uniparental applications would be informative. For example

in a species where female heterogametic sex determination system is operating (WZ/ZZ) all male stocks can be produced by applying one round of androgenesis from species of ZZ males. Each viable mitotic androgenetic is a singular isogenic individual with a unique genotype derived from a singular sire haplotype.

Alternatively, viable androgenetics can be produced using diploid sperm of a tetraploid parental fish by avoiding diplodisation process. However these techniques do not produce an isogenic progeny and most times involves hybridisation (Pandiran & Koteeswaran, 1998). Sun et al (2007) produced interspecific androgenesis using diploid sperm from allotetraploid hybrids of common carp × red crucian carp, more recently Zhou et al. (2015) reported the first time production of a viable diploid homozygous YY fish with unreduced diploid sperm of the autotetraploid fish, which were derived from distant hybridization. Induction of such androgenetic progeny through diploid sperm is of interest for genetic research and breeding purpose where significantly increased survival is achieved.

Similar to mitotic gynogenesis, mitotic androgenesis does not give rise to many viable fish. This is because in both techniques the shock treatment targets the first cleavage (optimal timing is difficult as the timing may vary among batches) and since only one set of maternal or paternal (in the case of mitotic gynogenesis and mitotic androgenesis, respectively) genetic material is being duplicated, every allele that reduces viability is expressed in the homozygous state. This significantly increases the mortality yet those that survive are free from major recessive deleterious alleles. Bertotto et al. (2005) observed a significant deviation from Hardy-Weinberg equilibrium in an experimentally produced progeny of mitotic gynogenetic in European seabass and suggested that those markers with significantly lower allele frequencies probably resulted from linkage to a

deleterious gene. Similarly Komen et al. (1992b) concluded that higher survival observed in meiotic gynogenetics was due to a masking effect of recessive deleterious alleles.

Androgenesis has been proven to be successful in many aquatic species (e.g. Bongers et al., 1998; Babiak et al., 2002; Patton et al., 2007) with limitations in some others, such as European seabass as a representative of marine species with small egg size (Colléter et al., 2014). The utility of androgenic progeny have been proven on (i) production of genetically isogenic inbred clonal lines, (ii) on the production of viable (YY) supermales in species with male-heterogametic chromosomes and (iii) conservation of sperm for gene banking (Babiak et al., 2002; Robison & Thorgaard, 2011). Although androgenesis requires strong gamma irradiations to inactivate the maternal nuclear genome, the mitochondria and associated mtDNA within the eggs are found to be unaffected (May & Grewe, 1993; Brown and Thorgaard, 2002). This creates an interesting field of study for investigation of several egg sources to study maternal effects in a genetically identical background. Brown et al. (2006) studied the effect of the mitochondrial genome on development rate and oxygen consumption in androgenetically produced rainbow trout and concluded that this had a significant role on early development rate among the clonal lines of rainbow trout. Therefore these authors suggested selection for mitochondrial genomes could increase growth rates and possibly food conversion ratios in aquaculture species.

Recently, an alternative to irradiation of the egg nucleus has been demonstrated by using a cold shock. Briefly, activated eggs are immediately shocked following fertilisation which results in over 30% of haploid embryos by eliminating the maternally derived genome, accompanied by a heat shock in the first mitosis leading to viable diploid androgenetic progeny. Although several studies presented results of using this cold shock approach for production of viable diploid androgenetic progeny in loach (*Misgurnus*

anguillicaudatus), zebrafish and Japanese flounder (*Paralichthys olivaceus*) (Hou et al., 2014, 2015, 2016; Morishima et al., 2011), very little has been explained about exactly how this technique works.

1.4 Isogenic clonal lines in fish research

Clonal lines are of considerable interest for aquaculture-related research. Their standardised genetic background simplifies the analysis of complex genetic traits just as in inbred mice strains. This makes them a unique tool for fish research in a wide variety of fields including reproductive biology, quantitative genetics, physiology, fish behaviour, nutrition, ecotoxicology and many more (Dunham 2004; Komen & Thorgaard, 2007; Robison & Thorgaard 2011).

Isogenic clonal fish lines have been produced in several fresh water and marine species as the most recent review dating back to almost one decade ago by Komen & Thorgaard (2007). Table 1.1 demonstrates the species where isogenic clonal fish lines have been successfully produced with only two extra studies (Hou et al., 2015, androgenetic zebrafish clonal lines and Liu et al., 2011, gynogenetics Japanese flounder) since 2007. This suggests limitations observed during production and maintenance of such lines, which will be discussed in section 1.4.5.

Table 1.1: Summary of isogenic clonal fish lines produced.

Common name	Species name	Type	Irradiation	Reference
Zebrafish	<i>Danio rerio</i>	Gynogenesis	UV	Streisinger et al. (1981)
		Gynogenesis	UV	Mizgireuv & Revskoy (2006)
		Androgenesis	Cold shock	Hou et al. (2015)
Medaka	<i>Oryzias latipes</i>	Gynogenesis	UV	Naruse et al. (1985)
Common carp	<i>Cyprinus carpio</i>	Gynogenesis	UV	Komen et al. (1991:1993)
		Androgenesis	UV	Bongers et al. (1997a)
		Gynogenesis	UV	Ben-Dom et al. (2001)
Ayu	<i>Plecoglossus altivelis</i>	Gynogenesis	UV	Hans et al. (1991)
		Gynogenesis	UV	Taniguchi et al. (1996)
Nile tilapia	<i>Oreochromis niloticus</i>	Gynogenesis	UV	Sarder et al. (1999)
		Gynogenesis	UV	Muller-Belecke and Horstgen-Schwark (1995)
		Gynogenesis	UV	Hussain et al. (1993)
Amogo salmon	<i>Oncorhynchus rhodurus</i>	Gynogenesis	UV	Kobayashi et al. (1994)
		Gynogenesis	UV	Qin et al. (2002)
Red seabream	<i>Pagrus major</i>	Gynogenesis	UV	Kato et al. (2002)
Rainbow trout	<i>Oncorhynchus mykiss</i>	Gynogenesis	UV	Quillet et al. (2007)
		Androgenesis	UV	Robison et al. (1999)
		Androgenesis	UV	Young et al. (1996)
Japanese flounder	<i>Paralichthys olivaceus</i>	Gynogenesis	UV	Hara et al. (1993)
		Gynogenesis	UV	Liu et al. (2011)

Production of isogenic clonal lines can be attained in two consecutive generations through mitotic gynogenesis and/or androgenesis. The methodology of choice depends on phenotypic sex (i.e: androgenesis from an XY male or gynogenesis from a WZ female would give rise to both sex progenies in A1/G1 is induced respectively in male and female so as to develop clonal lines in the next generation). In the first generation fully homozygous individuals can be produced (both G1 and A1 fish are 100% inbred). Once the first generation fish reaches maturity, gametes from such fish is used for a second round of gynogenesis or androgenesis so as to produce populations of genetically identical fish, each being a *clone* of the others in the same line and derived from a specific inbred clone founder. Figure 1.3 explains the production of isogenic clonal fish lines through mitotic gynogenesis in female homogametic systems with explanations throughout. In an effort to increase survival in resultant isogenic clone populations,

alternatively meiotic gynogenetics can be used with higher levels of survival, once G1 fish are successfully established with completely homozygous genome, during the second generation of producing clones from each clone founder. Although there will be recombination as all loci are homozygous this will not induce any heterozygosity in the offspring from a single parent.

Alternatively, chromosome set manipulations can be applied in species with sexual dimorphism in such a way that the earlier maturing sex can be used for faster development of isogenic lines. For example, in European seabass males mature earlier than females, thus androgenesis offers much faster production of isogenic clonal lines. However Colléter et al. (2014) indicated negative results on inactivating maternal genome through androgenesis regardless of the wide range of UV doses that were tested. Out of 76 putative androgenetic progeny derived from three families only one single larva showed fully paternal inheritance in one microsatellite locus (*Dla-22*) out of 9 diagnostic loci, while rest of the putative androgenetic progeny represented bi-parental contribution regardless of UV irradiation applied to eggs prior to fertilisation (some of the putative androgenetic larva represented only maternal inheritance - results derived from microsats were confirmed with ddRADseq analysis carried out later, results not shown). These results were suggested to indicate a more specific problem related to marine species, those with small pelagic eggs as in European seabass. Alternative methods offering replacement of UV by cold shock as successfully applied in zebrafish (Hou et al., 2015) and loach (Morishima et al., 2011; Hou et al., 2013, 2014) and Japanese flounder (Hou et al., 2016) may provide new insights into species reported ineffective UV irradiation treatment. In such a species with prime interest to aquaculture production, establishing isogenic lines can be beneficial as a resource for aquaculture-related research, for example defining sex determination QTL and interactions with the environment.

Palaiokostas et al. (2015b) recently provided an additional and more comprehensive support to the polygenic sex determination hypothesis in European seabass which was previously suggested by many researchers (Blázquez et al., 1998; Saillant et al., 2002; Vandeputte et al., 2007).

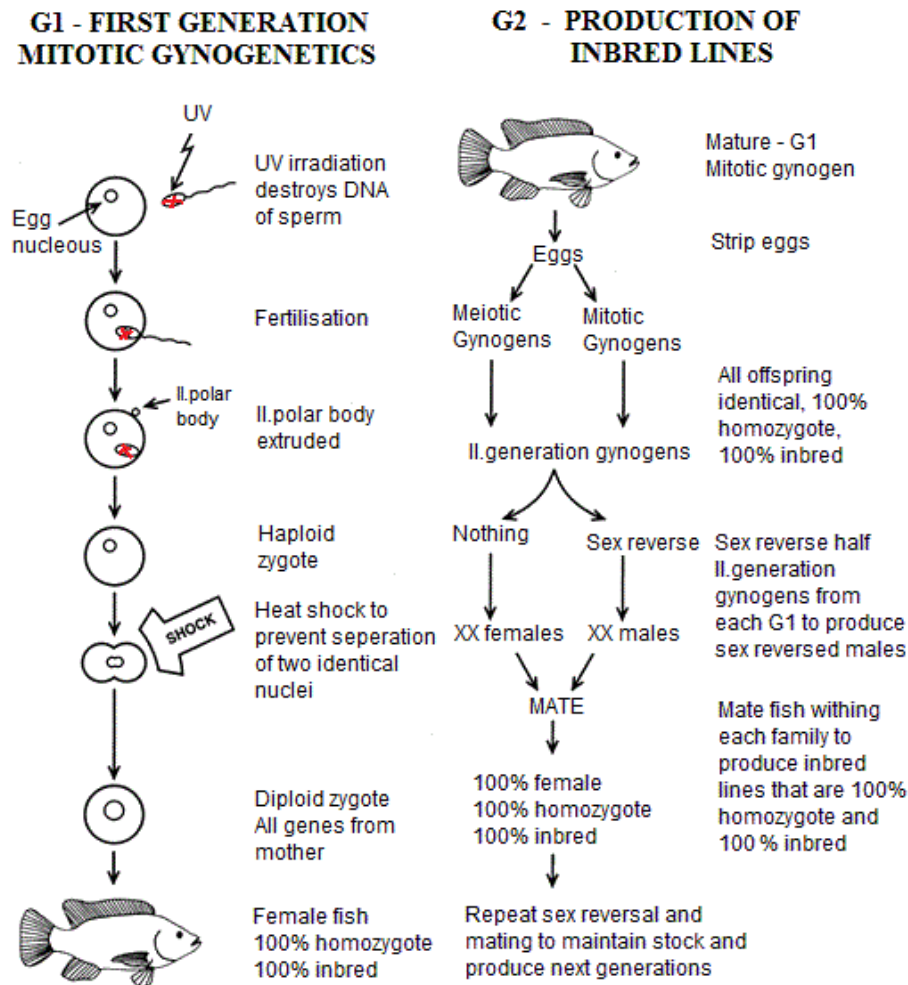


Figure 1.3: A schematic representation of the production of isogenic clonal fish lines in two subsequent generations in female homogametic species (XX/XY) through mitotic gynogenesis.

1.4.1 Sex determination

There are several ways of elucidating sex determination in the diverse group of fish order such as through cytological studies, by examining the sex ratio of progeny from sex-

reversed individuals and developing mono-sex populations, reviewed by Devlin & Nagahama (2002) and Penman & Piferrer (2008). However such techniques can be time-consuming depending on the reproductive maturity time in each species of interest. Uniparental reproduction techniques, on the other hand, enables fast identification of sex determination in any species of interest (which is the baseline to construct any further study and used to be a methodology of choice until recently) on sex determination and differentiation. However with the advances achieved in Genotyping by Sequencing (GBS) the focus is shifting toward to identify thousands of markers from one or more families and perform association studies so as to locate sex determining region (Palaiokostas et al., 2013a, 2015a) Mitotic gynogenesis and androgenetics offers the possibility of gaining insights into the sex determination system. In species where females are homogametic (XX/XY sex system), application of gynogenesis is expected to result in all-female progeny, while induced androgenesis is expected to result in equal proportions of females and males. YY individuals propagated via androgenesis in male heterogametic/female homogametic systems are viable in several fish species and these are of interest for production of mono-sex male populations once crossed with normal XX females. In the species where females possess heterogametic sex chromosomes (WZ/ZZ), the opposite is expected, where application of androgenesis is expected to give rise to a resultant progeny of all-males while induced gynogenesis in such family is expected to result in equal portions of male and female, unless the sex determining locus is located close to telomeric regions with undifferentiated sex chromosomes, as commonly encountered in fish, then the ratio would be 1:1 WW (if viable) : ZZ in DH mitotic gynogenetic while almost all WZ (female) in meiotic gynogenetic group. Similar to the previous scenario, WW females are viable in some fish and are of interest for the production of mono-sex female populations (in some species) once crossed with ZZ

normal males in species of female heterogametic sex systems (WZ/ZZ). There are records of obtaining both females and males in resultant progeny of gynogenetics in literature, as clearly shown in XX/XY sex chromosomal system of coho salmon (Piferrer et al., 1994), rainbow trout (Quillet et al., 2002) and in common carp (Komen et al., 1992a). These highlight uncertain types of sex determination system under operation and/or epigenetic factors or recessive mutations taking place in some fish species (Devlin & Nagahama, 2002).

Alternatively significantly lower survival of mitotic gynogenetics and androgenetics can be improved by inducing meiotic gynogenetics to elucidate how sex chromosomes are structured. In an effort to identify sex determination system in ship sturgeon (*Acipenser nudiiventris*) a family of experimentally induced meiotic gynogenetics were produced and results based on female biased sex ratio suggested the existence of female heterogametic sex determination system under operation (Hassanzadeh Saber & Hallajian, 2013). However sturgeon sex determination system is known to be very complex, given one example of meiotic gynogenesis might reflect only a part of the entire sex determination puzzle. Similarly Quillet et al. (2002) investigated a possible mutation (*mal*) in sex determining gene so as to understand fortuitously observed unexpected males in a mitotic gynogenetic family of rainbow trout (e.g: 13 males out of 27 survivors). Although results derived from three generations of conventional and/or meiotic and mitotic gynogenetics were not persuasive to draw any conclusion on the relative position of the *mal* locus, genetic analysis carried out in meiotic gynogenetic progeny revealed that the primary sex determining locus was located very close to the centromere. An example of such techniques with a direct industry application involves production of meiotic gynogenetic silver barb (*Puntius gonionotus*), with populations of all-female progeny being sex reversed in large numbers as broodstock for commercial production of mono-sex female

production to avoid the time-consuming process of progeny testing (Pongthana et al., 1999).

1.4.2 Linkage and gene-centromere mapping

Each type of uniparental reproduction technique offers a valuable genotype for genetic mapping. For example, mitotic gynogenesis and androgenesis represent unique genotypes with each derived from a singular parental source (maternal and paternal, respectively) in highly repeatable systems (assuming parental source is available). As doubled haploids share the same set of genome (in duplicated form, “2n”, as opposed to “n” in haploids) with haploids that are derived from the same uniparental reproduction technique, these can be used for replacement of one other. Kocher et al. (1998) produced the first linkage map in Nile tilapia based on 41 gynogenetic haploids. Alternatively, disadvantage of isogenic G1 and A1 progeny, due to reduced survival compared to backcross families, can be overcome by production of F1 hybrids then a subsequent generation (F2 or backcross) where two genetically distinct isogenic lines are crossed to produce progeny of heterozygous individuals with segregating informative alleles, providing more appropriate mapping panels. Young et al. (1998) constructed a genetic linkage map based on an androgenetically produced F1 hybrid progeny. Segregation analysis anchored 476 markers into 31 major linkage groups. The sex-determining locus was also located as a phenotypic trait, to a distal position in LG1 with the closest AFLP marker being located at 19 cM apart from the locus on the genome of rainbow trout. This map was improved later on by integrating a half-tetrad analysis using a meiotic gynogenetic family to locate 15 centromeric regions which were previously not well represented (Sakamoto et al., 2000). Shortly after Nichols et al. (2003) utilised a doubled haploid family of androgenetics produced in rainbow trout so as to update and consolidate the existing linkage maps from

previous studies (May & Johnson 1990; Young et al. 1998; Sakamoto et al. 2000). More than 900 markers were added to the consolidated map with a total density of 1359 markers. This map not only constituted the most comprehensive map, with higher density, but also enabled further examination of complex traits in rainbow trout by dissecting the QTL component influencing time of hatching in rainbow trout. Afterwards, another microsatellite map covering a whole set of chromosome arms (52) was generated by Guyomard et al. (2006). This map also incorporated data from a meiotic gynogenetic family so as to locate centromeric regions which was in accordance with fluorescent in situ hybridization results. In another study, a dense genetic linkage map genome based on gynogenetic haploids revealed strong chiasma interference in sockeye salmon, a well-known phenomenon in fish species. The same map also incorporated a meiotic gynogenetic family to locate centromeres (Limborg et al., 2015).

The way meiotic gynogenetics are produced, by retention of second polar body, arrests the result of any cross over during the second meiosis. It is due to the fact that sister chromatids (attached at the centromere) do not separate until meiosis II is completed. This feature of meiotic gynogenetic progeny provides an irreplaceable tool for identifying recombination of a locus with its centromere along the length of the chromosome and in the whole genome of interest. These sister chromatids tend to be homozygous with respect to genes near the centromere and heterozygous for genes further away from the centromere due to difficulties of having a crossing over in shorter distances. For example in the absence of recombination, sister chromatids carrying identical DNA fragments will be homozygote whereas in the case of recombination (genetic material exchange) sister chromatids will be heterozygote. Hence, simple scoring of a meiotic gynogenetic full-sib progeny allows positioning of a locus with respect to its centromere (Danzmann & Gharbi, 2001).

Although recombination events are expected to be random along the chromosomes (Guo & Allen, 1996), significant differences are observed both along the length of chromosome as some regions of the chromosomes are being more active than the others and between sexes (Komen & Thorgaard, 2007). Centromeres are genetically the least active parts of the chromosomes, governing proper chromosome segregation during cell division. Since such regions are functionally playing a significant role, any adjacent locus is less likely to recombine. As the distance of a given locus increases along the chromosome from its centromere the more likely it is to be heterozygous in a meiotic gynogenetic. Regarding to the marked differences observed between sexes, the Salmonidae family is the best known example where most females experiences more recombination events and the distribution of crossover events are more homogenous as opposed to telomere-specific recombination patterns observed in males. Atlantic salmon represent the largest female:male recombination ratio difference (i.e., 16.81:1) compared with rainbow trout (4.31:1) and Arctic charr (1.69:1) (Danzmann et al., 2005).

Gene-centromere mapping through half-tetrad analysis offers direct value on orienting mapping data into a framework that may be related to a physical map of the genome. Such unique characteristic of meiotic gynogenetic has widely been exploited in fish research. One of the earliest works, carried out by Thorgaard (1983) using a family of meiotic gynogenetics, revealed high interference over long map distances in rainbow trout. Later gene-centromere maps have been developed in several fish and shellfish species (Martínez et al., 2008; Nie et al., 2011, 2012; Zhu et al., 2013).

1.4.3 QTL mapping

Quantitative traits are mostly consequences of multiple genes and environmental factors influence the expression of the phenotype. Such traits define the sections of DNA that are

associated with variation observed in a phenotype of interest. Analysis of such traits is a complex task. However, inbred lines with reduced background noise are ideal for this complex task (Balding et al., 2007).

Genetic linkage maps are the baselines of QTL studies. Detection of a major QTL with high power and precision in an inbred line cross depends on the genetic diversity between the parental strains, the heritability of the trait under investigation, the size of the mapping family and the density of the genetic markers involved in the study. Therefore the prerequisite of any QTL study is to start off with as genetically divergent stock as possible (e.g. resistant and susceptible lines are used for the investigation of disease resistance). The first step towards such analysis is to identify phenotypic variation. Some quantitative traits including body length, body weight and meristic counts of isogenic clones produced by mitotic gynogenesis in ayu (*Plecoglossus altivelis*) were identified (Taniguchi et al., 1994). Likewise Robison et al. (1999) examined the genetic basis of developmental rate by using clonal lines of rainbow trout and identified significant differences in the physiological time for hatching. Hence, data indicated a strong genetic component in this trait and showed the suitability of such trait for QTL identification later on. Quillet et al. (2007) investigated disease resistance in nine isogenic clonal lines of rainbow trout from a domestic population and observed high variability between clones for resistance to the viral haemorrhagic septicaemia virus (VHSV). Overall three clones were highly resistant to VHSV, with over 95% survival in repeated experiments, while others were highly susceptible. Survival in the original population was 16% while 0-99% survival observed in different clonal lines, suggesting that the cloning process has fixed different genetic responses to VHSV. Clonal lines with such extreme phenotypes (resistant and susceptible) are valuable tools for investigation of the genetic components (QTL, candidate genes) involved in disease resistance in rainbow trout. Lucas et al.

(2004) investigated the variation in behaviour patterns by propagating four clonal lines of rainbow trout and reported a significant genetic effect on swim level, hiding, foraging, startle response to a sudden threat and aggression level among clonal groups. Since identical experiments were undertaken in identical conditions in common garden experiments, the results suggested strong genetic differences among clonal lines. Two clonal lines recently derived from populations reared in captivity for over a hundred years exhibited reduction in predator avoidance patterns and increased aggression compared to progeny of two clonal lines from more recently domesticated populations (Lucas et al., 2004). Detailed further investigations on the basis of identifying genetic factors of such differences could provide insightful information on behavioural patterns influenced by domestication, which would be of direct interest for aquaculture point of view. Similarly Millot et al. (2014) used seven heterozygous isogenic lines of rainbow trout so as to investigate fish personality traits by challenging each group of clones to a range of experimental situations (such as risk taking and fight response towards to a stressor). These clones provided a unique tool for investigation of inter-individual variability of fish personality and helped establishing phenotypes of low and high responsive groups in each stimulus. These are of interest for future QTL detection. Studies as such rarely take place due to the difficulty in controlling fish genetic origin and life history. To this end the research of Millot et al. (2014) was important to progress in this area. Zimmerman et al. (2005) reported a single major QTL controls natural killer activity cell-like involves in IPN resistance in a rainbow trout genome derived from a hybrid of two divergent clonal lines via androgenesis. In an effort to understand the genetic basis of smoltification-related traits such as growth and condition factor, body coloration, morphology, and osmoregulatory enzymes during the smoltification period in rainbow trout, Nichols et al (2008) made a genetic cross of clonal lines derived from migratory and non-migratory

life-history types in this species. The results were not conclusive with several genomic regions are being associated with smoltification or the physiological and morphological transition that occurs prior to seaward migration. However, parallel studies as such can unravel evolution of anadromy in Salmonidae, and thus increase our understanding on how the genetic component influenced by environment factors affects the smoltification process in salmonids.

The standardised genetic structure of isogenic clonal lines with complete homozygosity offers the possibility of designing mapping crosses where the segregation of only one parent can simply be traced. Palaiokostas et al. (2013b) used two families derived from an isogenic clonal line (female) and outbred XY males to map the major-sex QTL in LG1 in Nile tilapia in fine detail. Similarly Ozaki et al. (2001) used a mapping family panel propagated from a backcross between an IPN-resistance / susceptible strains of rainbow trout and identified two QTLs affecting disease resistance suggestion it is a polygenic trait. However the existing of one major QTL effecting IPN resistance was well documented in upcoming years (Houston et al., 2008, 2012; Moen et al., 2009). This constitutes a recent breakthrough application of genomics applied to aquaculture production and breeding.

The simplified nature of QTL identification in inbred lines has widely been acknowledged in species such as rodents, used as a model for human disease with direct medical application (Li et al., 2005), and in barley (Chutimanitsakun et al., 2011). Research involving isogenic clonal lines has reached to utilise recombinant inbred lines (RIL) in plants. Such lines are produced by crossing parental strains to inbred strains as a source for extensive mapping (Bertioli et al., 2014). These RIL have been extensively characterized for numerous important phenotypes which accelerate studies on the basis of complex traits (Klasen et al., 2012). A similar approach to RIL has been used in mice

research as a model for mammalian research with direct application to human health (Zou et al., 2005). However QTL mapping studies based on mapping panel of isogenic lines in all aquatic species lags behind: many maps until recently were produced predominantly from mapping panels of outcrossed individuals.

1.4.4 Genetic and genomic resources

Emerging genomics technologies have increasingly been enabling whole genome sequencing projects to be carried out in aquatic species. However this is not an easy task to perform in fish genomes (i) due to large genome sizes in some fish species (in some species such as Atlantic salmon, the genome size is as large as the human genome, 3.4×10^9 bp, while in some other species such as fugu, as a model species for research, the genome size is one of the smallest found in vertebrates at 4×10^6 bp) and (ii) a fish-specific whole genome duplication event (4R-WGD) that took place in the course of evolution (see section 1.6) complicates assembly. As of today, even with the availability of high-throughput sequencing technologies (see section 1.5), assembling the genomes of aquatic species is challenging and an approach to help overcome such complications is to sequence a completely homozygous genome of doubled haploid individuals (Davidson et al., 2010). This fruitful approach has been adopted to reduce ambiguity in the assembly procedure in several aquatic species (Howe et al., 2013; Brawand et al., 2014; Berthelot et al., 2014; Xu et al., 2014; Lien et al., 2016) and will clearly be a methodology of choice for future whole genome sequencing projects. For example, in Nile tilapia the first genome assembly (Brawand et al., 2014) covered about 70% of the total genome. Recent improvement in the assembly (unpublished: TD Kocher, pers. comm.) was achieved when data from a dense genetic linkage map with over 3700 SNP markers (Palaiokostas et al., 2013b) and long read PacBio sequencing (Carneiro et al., 2012) were combined with the

previous data. Benefits of utilising long read sequencing have been reported in puffer fish, *Takifugu rubripes*, where the genome assembly was dramatically improved by utilising doubled haploid specimens (with almost no allelic differences in genome) in comparison with wild type counterparts (Zhang et al., 2014). This is due to DNA polymorphism observed in outbred organisms making it difficult to determine the origin of real sequence differences, whether due to allelic polymorphism or among the repeat regions of a genome. PacBio long read sequencing, as in most NGS data analysis, stacks all identical reads (lower read depth compared to short read sequencing of Illumina platforms) one on top of each other. Using DH genomes significantly reduces the chances of mistaking artefactual (sequencing error) variation for real variation, therefore making the assembly procedure easier. Pacbio long read sequencing developers advise to select strains (such as inbred lines or DH strains) to minimise heterozygosity which appears as separate contigs (see tutorials in experimental design on PacBio official website). Having access to an assembled genome of a species of interest can lead to better management of stocks with a direct interest to aquaculture.

Furthermore such lines should improve the accuracy of microarray and/or RNAseq studies where genetic variation exists in the outbred source that can be a major source of variation. Therefore small numbers of biological replicates are of critical interest for the normalisation of the expression data. This is where isogenic lines come in with genetically uniform structure. Utilising homozygous genetic background of such clones and/or inbred individuals provide an important tool where variation within each line is minimised among biological replicates while variation among lines is maximised compared to an outbred source in a species of interest. Hence this approach results in improving the detection of differentially expressed genes between lines with much higher statistical power and accuracy. The validity of such an approach has successfully been

demonstrated using inbred lines of mice (Wei et al 2004). Three androgenically derived rainbow trout clonal lines used for transcriptome profiling from both liver and head kidney tissues revealed differences among three allopatric populations. Variance in clonal lines was statistically not significant while variance among the different clonal lines exhibited constitutively different transcript abundance for a subset of genes diverged from three allopatric populations (Bayne et al., 2006). Although this study did not take into account the limited sampling that each clonal lines offer (section 1.4.5). Since each clonal line is derived from a single parent exploiting such information to differentiate stocks from different allopatric populations may be beyond findings, considering representation of source population can be limited in clonal lines. Purcell et al. (2006) carried out a comprehensive gene expression profiling following DNA vaccination of homozygous rainbow trout against hematopoietic necrosis virus (IHNV) and study revealed that the IHNV DNA vaccine induces up-regulation of the type I IFN system across multiple tissues, which is the functional basis of early anti-viral immunity.

1.4.5 Limitations of isogenic clonal lines

As in most scientific techniques used in biological organisms, isogenic clonal fish lines production has pitfalls which can be classified as: (i) residual chromosome fragments from irradiated gamete source; (ii) poor survival rates; (iii) poor fertility rates of DH fish and (iv) spontaneous arise of meiotic gynogenetics in DH mitotic gynogenetic group.

The first limitation, residual chromosome fragments from the irradiated gamete source, is due to suboptimal irradiation, which requires optimisation per species. Various irradiation techniques have been used including X-ray, UV and gamma irradiation (Thorgaard 1983; Pandian & Koteeswaran 1998; Overturf 2009) however, the choice of technique is depending on the gametes to be irradiated as well as practicality and safety

considerations. Although both gamma and X-ray offers higher penetration power, UV has been the methodology of choice for induction of both gynogenesis and androgenesis in fish due to its availability (Komen & Thorgaard, 2007). Foisil & Chourrout (1992) demonstrated that the yield of gynogenetic fry obtained with UV irradiated sperm was higher than that of obtained with gamma irradiated sperm. UV irradiation is mostly effective in diluted layers of milt and eggs (Chourrout et al., 1986). As a general procedure, the optimal irradiation dose is identified by first using a wide range of different irradiation doses and defining the most efficient dose leading to haploids based on their highest survival rate. Suboptimal doses may result in persistence of fragments of the irradiated genome in the resultant experimental progeny while high doses of irradiation lead to reduced viability of the irradiated gametes and subsequently poor fertilisation and survival. Only optimal dose of irradiation ensures complete inactivation of the genome of irradiated gametes but still leads to relatively high rates of fertilisation. This is a common phenomenon in both gynogenesis and androgenesis procedures. For example, in the process of optimising androgenesis high irradiation doses applied to fish eggs can lead conflictingly low survival due to damage that is caused on maternal mRNA which is essential for the first phase of embryonic development (Pelegri, 2003) as well as fragmentation of the maternal DNA (Colléter et al., 2014; Ocalewicz et al., 2012). In contrast, an insufficient dose of irradiation to eggs results in insufficient inactivation of the maternal nuclear genome which triggers large chromosomal fragments to be retained in the resultant androgenic progeny. Similar effect is observed in irradiation of sperm during induced gynogenesis: higher doses of irradiation applied to spermatozoa lead to motility lose therefore fertilisation fails or lower doses of irradiation used causes large numbers of residual chromosome fragments in the resultant progeny due to insufficient irradiation doses which may impair embryo viability (Chourrout and Quillet, 1982).

Literature has reports of chromosome fragments even in the optimised protocols (Ocalewicz et al., 2004). Therefore optimised protocols for the inactivation of gametes are one key prerequisite for the development of isogenic clonal fish lines.

A second limitation hampering successful production of isogenic clonal fish lines is poor survival observed in mitotic gynogenetic and androgenetic fish, which can be due to technical and biological reasons. Technical reasons involve the application of physical and mechanical influences to gametes which reduces survival. The more external challenges that gametes are exposed to, with increasing handling time, the lower the survival will be. The effect of the temperature or pressure shock used to restore diploidy has been suggested to affect a range of other mechanisms involved in early embryo development. Both pressure and heat shock in gold fish and crucian carp induced developmental disorders including dorsal deformities in embryos and reduced survival (Yamaha et al., 2002). As both mitotic gynogenesis and androgenesis result in completely inbred individuals, there is no heterozygosity to mask any recessive deleterious alleles. The double expression of such alleles results in significantly reduced survival in clone founder progeny. This in conjunction with physical shock applied to restore diploid reduces survival markedly in the production of DH fish. One point that needs to be addressed is that epigenetic programming in the early developmental stages plays an important role in survival of uniparental individuals. This has been well studied and understood to some extent in mammals, however still remains to be investigated among other vertebrates. Potok et al. (2013) characterised the genome-wide pattern of DNA methylation profile in zebrafish at various stages and found that the sperm methylation profile was remarkably similar to the profile of the zygote while the oocyte methylation profile was distinct from both the sperm and the zygote at the early developmental stage. The same authors further validated the function of the sperm methylome pattern by

inducing mitotic gynogenesis (using UV-inactivated sperm) and reported that a competent paternal genome is not required for DNA reprogramming of the zygote. In other words, methylation pattern of early developmental stage zebrafish is determined by the sperm, even in gynogenetic group after UV inactivation of the sperm. Similarly, Jiang et al. (2013) provided further evidences on DNA methylome is inherited by sperm as opposed to oocyte in early embryos of zebrafish by performing both gynogenesis and androgenesis. The biological reasons for low survival may include variation in early egg development rate among batches from the same female. It may thus be hard to achieve precise shock timing. Hence some authors applied a range of shocks as in Chapter 6 (this thesis) in meiotic Nile tilapia. Francescon et al. (2004) adjusted the time of shocking among different females by producing small numbers of bi-parental controls to observe the first cleavage furrow. These authors indicated European seabass could show first cleavage furrow with a difference of up to 15 minutes (Bertotto et al., 2005). The yield obtained from gynogenesis and androgenesis was similar among the studies, ranging between a minimum of 1% in African catfish gynogenetics (Galbusera et al., 2000) to a maximum of 23% as reported in rainbow trout gynogenetics (Diter et al., 1993). Nam et al. (2002) reported 19% survival at hatching from interspecific androgenesis between heterozygous mug loach sperm and UV irradiated common carp eggs. More recently Kucharczyk et al. (2014) reported over 2% survival for haploid androgenetic common tench (*Tinca tinca*) at hatching stage while lower survival was detected in DH androgenetic common bream (*Abramis brama* L.). The final yield of DH fry at first-feeding stage varied between 5 to 10 % in common carp (Komen et al., 1991; Bongers et al., 1994), one of the most successfully studies in fish species on isogenic clonal lines production. There are also records of failed attempts in literature which are as significant as successful application of

DH fish (i.e: androgenesis reported in muskellunge by Lin & Dabrowski, (1998) and androgenesis reported in European seabass by Colléter et al. (2014).

A third limitation hindering successful production of isogenic clonal fish lines is poor fertility rates observed in DH fish. Sterility reported in the literature is quite diverse ranging from 13% to over 90% detected in androgenetic common carp. In an effort to elucidate the sex determination system in common carp, Bongers et al. (1999) produced homozygous YY individuals, however high numbers of sterile fish were observed (70-94 % in androgenetic group; 37-55% in controls), hampering the analysis. However high numbers of sterile individuals in both control and androgenetic group could not be explained. In total of 77 gynogenetic DH fish produced in Nile tilapia, 10 produced viable eggs (Müller-Belecke & Hörstgen-Schwark, 1995). Quillet (1994) studied reproductive traits of mitotic gynogenetic rainbow trout and reported that DH females showed a week delayed spawning activity with longer reproductive season due to variability observed in females. Absolute fecundity of DH females was significantly lower than that of controls however relative fecundity (per kilo body weight) was almost equal to diploid controls. In general, female progeny derived from mitotic gynogenesis and androgenesis represented poorer survival (e.g: reduced egg size and quality, decreased ovulation rate) than that of DH male androgenetic progeny: this is thought to be due to the more complex nature of female reproduction (Komen & Thorgaard, 2007).

Last but not the least, each clonal line represents a very narrow sampling of the total genetic variation of a species, since they are initially derived from a single DH female or male. Therefore, depending on the objectives of the experiment, the best strategy is to use a small number of individuals (since they are clones with identical genetic background response should be similar) derived from multiple clone founders, thus effect of many genotypes can be estimated with higher accuracy (Festing & Altman, 2002).

1.5 Duplicated fish genomes and their complications

Scientific evidence suggests that two successive round of whole genome duplication occurred in the ancestors of vertebrates: one before and one after the divergence of the lamprey (jawless fish, order Petromyzontiformes) lineage (500-800 mya) (Wolfe, 2001). The first of these WGD separated cephalochordates from early agnathans (jawless fish) in the course of evolution (Fig 1.4). A third round of WGD took place 300-400 mya in the lineage of teleosts following their divergence from basal ray-finned non-teleost fish, termed as the teleost-specific WGD (Ohno, 1970; Opazo et al 2013; Glasauer & Neuhauss, 2014). Some fish species, namely salmonids (Allendorf and Thorgaard, 1984; Allendorf et al., 2015), common carp (Larhammar and Risinger, 1994) are believed to have an additional (4R) round of genome duplication much later in the course of evolution. This recent WGD event is thought to have provided genetic raw materials for the physiological, morphological and behavioural diversification of these species.

Gene and/or genome duplications generate new raw material for species to evolve by creating sufficient material to enable evolution through natural selection. It has been suggested that since WGD doubles the entire genomic material of a species, it must be significant for generating novel genes (Glasauer & Neuhauss, 2014). Studies confirmed that genome duplication has shaped the genome of eukaryotic organisms throughout evolution, and offered evidence to hypothesise its potential significance as a major evolutionary mechanism for speciation and diversification (Seoighe & Wolfe, 1999; David et al., 2003; Hoegg & Meyer, 2005). A WGD event can occur either through hybridisation between closely related species (termed as allopolyploidisation) or within a species (termed as autopolyploidisation) via failure of cell division by lack of disjunction among daughter chromosomes after DNA replication during meiosis (during gametic non-reduction) and mitosis (genome doubling) (Levasseur & Pontarotti, 2011).

Allopolyploids represent decent structural dissimilarity between their genomes to generate bivalent pairing during meiosis, hence represents disomic inheritance (David et al., 2003). In autopolyploids, however, originally identical chromosome sets pair up as multivalents at meiosis, which frequently result in unviable aneuploid gametes (Li, 1997). Thus, such changes in ploidy status are expected to be deleterious which serves as evolutionary dead-ends due to physiological or developmental constraints (Mable, 2004).

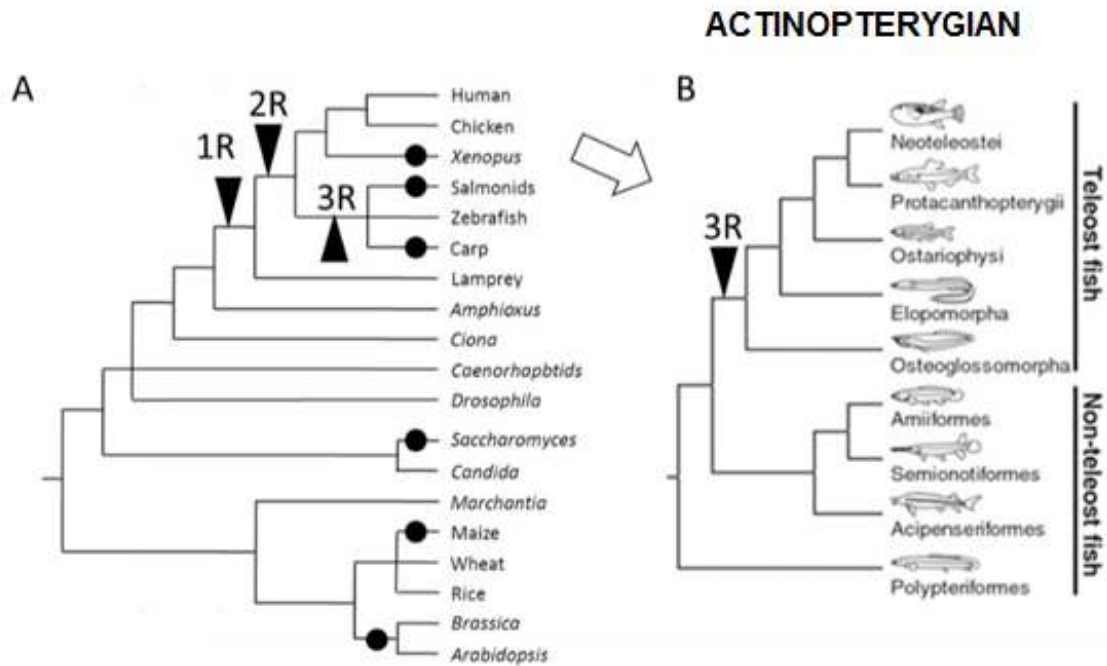


Figure 1.4: Whole-genome duplication (WGD) events during eukaryotic evolution. 1R, 2R, and 3R indicate first, second, and third-rounds of WGD in vertebrate evolution, respectively. A) Black circles indicate lineage specific WGD events, which took place in vertebrate, plant and yeast lineages. B) Actinopterygian phylogeny and the estimated timing of the 3R-WGD. Black circles indicate lineage specific WGD events (Adapted from Sato & Nishida, 2010).

According to Ohno's classical view the evolution of the genomes tend to be conservative, in the case of no duplication, in agreement with null hypothesis of *purifying selection* or *negative selection* by removing the alleles that are deleterious from the population. Therefore under the assumption of biological evolution is a progressive and gradual process, which results in optimal adaptive functions for survival of a specific population at a certain time and space; it is not wrong to consider that most extant species today are adapted to their current environments. Unless there is a change in the environment, the major mechanism of selection is towards to purifying selection, which maintains the adapted genetic function. Right after gene and/or genome duplication, duplicated genes are redundant and theoretically free to diverge at an accelerated rate due to relaxed purifying selection (Castillo-Davis et al., 2004). This, combined with the functional redundancy due to presence of two originally identical copies, the common fate of the

majority of duplicates is to be inactivated (nonfunctionalisation) via accumulation of deleterious mutations or disappear as a result of dynamic chromosomal rearrangements. However, such duplicated genes termed as *paralogs* can be retained under the strong negative selection in the case of extra protein (final product of coding DNA) being beneficial due to increased gene dosage effect or if they are diverged for a new adaptive, beneficial function (neofunctionalisation) or if ancestral function is partitioned or shared by both duplicates (subfunctionalisation) (Glasauer & Neuhauss, 2014). Therefore, until such duplicates diverge into their new function and/or disappear from the genome so that the duplicated genome reverts back to a stable diploid state. Such an evolutionary process presents complications in identifying true allelic variation from sequence variants that have been introduced to paralogous sequences in the course of evolution after WGD (See Chapter 5 of this thesis).

Although a WGD event provided raw material to be used for adaptation via natural selection, such events followed by differentiation of many gene duplicates via dynamic chromosomal rearrangements have caused one of the most complex animal genomes in ancestrally duplicated fish species such as Salmonidae family members (Danzmann et al., 2008).

1.6 Next Generation Sequencing (NGS) technologies

The advances in sequencing technologies were triggered with the announcement of Human Genome Project in 1990 and its estimated cost of \$3 billion from the initial stage up to the completion date of 2005. However the project was completed two years ahead of the scheduled timeframe with a total cost of \$2.7 billion dollars. This accelerated delivery time and reduced cost was due to facilitation of the very first stage of next generation sequencing technologies of the time, which could not be predicted at the

beginning of the project. Such high demand started off a whole new era of rapid, affordable and accurate genome analysis which was not only limited to humans but also used in many other organisms. While first generation sequencing technologies were considered to be the high-throughput of the time, the main acceleration was achieved when short read, massively parallel sequencing technologies were introduced to the market in 2007. Starting from that point onwards, the data output capacity of next generation sequencing technologies surpassed Moore's law – more than doubling each year, with a substantial decrease in sequencing cost (Liu, 2011). The Genome 10K consortium announced a plan to sequence the genome of 10,000 vertebrates (almost one representative for every vertebrate genus) so as to better understand the genomic functions of each species and for conservation purposes at realistic costs (Haussler et al., 2009). As predicted earlier (Mardis, 2006), a commercial scale bioinformatics company announced at 2014 to sequence human genomes at \$ 1,000, making it accessible to the public (Illumina HiSeq X Ten system with 30X coverage).

Regardless of several platforms available for NGS libraries, all require fragmentation of the target genome at an early stage. This can be achieved either using sonication (through random fragmentation) or through enzymatic fragmentation. Due to difficulty of optimising desired fragment size range with sonication and the lack of reproducing the same fragments made enzymatic fragmentation as a methodology of choice. DNA fragmentation in construction of NGS libraries utilises type II restriction enzymes. Such enzymes have the ability to cut the DNA from a fixed position with respect to their recognition site hence generates highly reproducible fragments that can be easily detected on gel electrophoretic systems. Briefly, library construction involves fragmentation of genomic DNA of interest using type II restriction enzyme(s) followed with ligation of synthetic, custom-made adaptors involves PCR primer binding site to be used during

genotyping by sequencing, restriction enzyme cut site and a specific set of molecular identifier fragments per sample. Following sequential purification steps to remove library from traces of enzymes, master mixes and co-factors in reaction mixes, size selection is carried out to select the desired size fragments for sequencing run (300-700 bp for short read sequencing in Miseq, Illumina). The last step prior to sequencing is the enrichment of the library via PCR followed with a final purification which eventually makes NGS library ready for sequencing.

Genotyping by sequencing takes place on a solid surface (e.g: bead) where covalently attached adaptors of bead hybridise with each library adaptor. A fundamental approach is to bridge amplify the library on a surface where massively parallel sequencing takes place as amplification continues (more fluorescent signals are produced). In NGS platforms everything happens in a step by step fashion where data production is followed by data detection of massively parallel sequencing yields. However one limitation of such platforms is the requirement for more bioinformatics work to be undertaken since the raw data output is massive and in smaller pieces (an average of 50-600 bp long fragments are generated).

NGS has been validated and widely applied in studies to identify population structure in yellowfin tuna (Pecoraro et al. 2016), for conservation genomics in European hake (Milano et al. 2011), to explore the larval transcriptome of the common sole as a candidate for European aquaculture (Ferraresso et al. 2013), for genetic linkage map construction in hāpuku (Brown et al. 2016) and QTL mapping in European seabass (Babbucci et al., 2016) , to elucidate epigenetics of fish sex differentiation (see review: Piferrer, 2013) and constructing genome assemblies for aquatic genomes by combining hybrid models so as to improve quality of the genome assembly (Lien et al., 2016).

1.6.1 Restriction based platforms (RADseq & ddRADseq)

Restriction based platforms are the easiest and cheapest way of generating large numbers of Single Nucleotide Polymorphism (SNPs), which are the most abundant form of genetic variation in eukaryotic genomes (Liu 2011). Restriction-site Associated DNA sequencing has been widely used since the method was published at 2008 (Baird et al., 2008). These authors identified three QTLs by generating over 13K polymorphic SNP markers in two populations of stickleback from a total of 96 individuals, as well as validating suitability of the technique for large scale genotyping and genetic mapping.

Later on Peterson et al., (2012) modified the RAD technique by using digestion of DNA by two restriction enzymes simultaneously, hence the name double-digest RADseq. This modification not only decreased the cost of sonication and end repair involved in RADseq procedure but also resulted in less DNA losses, thus library start gDNA concentration could be decreased down to 100 ng or less per sample.

Both RAD and ddRAD techniques create reduced representation of the genome by allowing over sequencing of the nucleotides next to restriction sites, hence detection of SNPs. Furthermore both provide strict control over fragments that are generated from massively parallel DNA sequencing across many individuals, thus allowing detection of SNPs within the magnitude of thousands (McCormack et al., 2012). However the main difference between these two techniques lies in the practicality of the procedure and fast library construction that can be applied in ddRADseq. The efficacy of both methods at generating large number of markers for exclusive genome scan has been validated by many studies (Davey et al., 2011; Recknagel et al., 2013; Gonen et al., 2014; Brown et al., 2016).

RADseq and its derivative ddRADseq have been widely exploited in fish genetics in studies of elucidating sex determining loci in Atlantic halibut, Nile tilapia and hāpuku

(Palaikostas et al., 2013a, 2013b; Brown et al., 2016), genetic linkage map construction from various mapping panels in Japanese eels and gudgeons (Cyprinidae) (Kakiora et al., 2013; Kai et al., 2014) for the purpose of population genomic studies in pearly oyster and stickleback (Catchen et al., 2013; Lal et al., 2016), QTL mapping in Atlantic salmon and European seabass (Houston et al., 2012; Palaikostas et al., 2015b), SNP chip development in Atlantic salmon (Houston et al., 2014) and understanding of the mechanisms of evolution in salmonids and spotted gar as an outgroup for teleost-specific whole genome duplication (4R) (Amores et al., 2011; Everett et al., 2012). Aquatic species genome projects also facilitated genetic linkage maps generated by RADseq and ddRADseq as a means to integrate genetic linkage map order with that of physical locations of the chromosomes (Howe et al., 2013; Brawand et al., 2014; Lien et al., 2016).

1.7 Aims and the objectives of the thesis

The aim of the research presented here was to gain insight into the development of isogenic clonal fish lines by addressing bottlenecks that have been encountered, using high-throughput sequencing technologies. To this end, the first focus was to address spontaneous meiotic gynogenetics in the production of isogenic clonal lines and identify highly informative telomeric markers to distinguish these from DH fish. This was then followed up with a genome-wide verification study in putative isogenic clonal lines; starting from outbred to G1 and to G2, so that reliable establishment of such lines could be proven by overcoming limitations of earlier marker technologies. One species with a duplicated genome (Atlantic salmon) was also included to investigate whether High Throughput Sequencing platforms, designed initially for non-duplicated genomes, present any complications in developing and applying such marker sets compared to those with non-duplicated genomes (European seabass and Nile tilapia).

The specific objectives of the current thesis were;

- i. To assess potential residual contribution from irradiated gametes, confirming uniparental inheritance (all experimental chapters; namely *Chapter 3, 4, 5 and 6*)
- ii. To explore genome-wide isogenicity in experimental groups (*Chapter 4 and Chapter 5*)
- iii. To generate a genetic linkage map based on meiotic gynogenetics (*Chapter 3 and Chapter 6*)
- iv. To locate centromeric regions by building a locus-centromere map (*Chapter 3*)
- v. To identify telomeric markers with higher recombination frequencies from those of centromeric markers so as to differentiate meiotic gynogenetics and mitotic gynogenetics (*Chapter 3 and Chapter 6*)
- vi. To validate the efficacy of HTS technologies by comparing with that of recent marker technology (microsatellites) in the case of observing false positives in putative doubled haploid mitotic gynogenetic progeny (*Chapter 4*)
- vii. To assess the power of well-established genomic resources of the Atlantic salmon on investigation of genome-wide isogenicity so as to remove multi-copy loci (*Chapter 5*)
- viii. To investigate whether HTS platforms represent complications working with duplicated genomes compared to non-duplicated genomes (*Chapter 5*)
- ix. Overall, to provide pilot studies on how HTS technologies can be applied to genome-wide verification studies so that the scale of the studies and the direction of the marker choices could be shifted towards to thousands of SNPs as opposed to handful of markers as used until recently. Hence this would ensure reliable establishment of isogenic clonal lines by detection of possible

residual chromosome fragments from irradiated gametes and false positives with genomically comprehensive marker technologies.

Chapter 2

General Material and Methods

2.1 General information

Three different species were used in the present study, Nile tilapia, Atlantic salmon and European sea bass. Handling of live fish, however, was only required for the Nile tilapia study which were held in the tropical aquarium of Institute of Aquaculture, Stirling, since the other fish were held by project partners in the established research facilities for each species (European seabass experiments were carried out at the Ifremer station in Palavas les Flots, France; Atlantic salmon experiments were carried out at the Institute of Marine Research (IMR), Bergen, Norway). Optimised production protocols in each species can be accessed online through the project website (<http://www.aquaexcel.eu/index.php/2016-02-15-20-04-23/deliverables>).

2.1 General maintenance of the Nile Tilapia stock in Tropical Aquarium

2.1.1 Fish stock origin and regulation

The Nile tilapia (*Oreochromis niloticus*) stock was introduced to the University of Stirling in 1979 from Lake Manzala, Egypt. The basic maintenance of the experimental stock in the tropical aquarium followed working procedures under Animals Scientific Procedures Act, 1986 (ASPA) and monitored by the Home Office (HO) in the United Kingdom. Based on HO guidelines, an accredited training must be performed for all personnel working under ASPA prior to carrying out any experimental work. Hence, all procedures of fish breeding including anaesthesia, tagging, sampling and applying chromosomal set manipulation were performed under project (PPL 60/1967) and personal (PIL 60/14087-IEC93D903) licenses issued by the UK Home Office.

2.1.2 General maintenance of stock

The temperature of the water in the Tropical Aquarium (TA) facility was maintained at 28 ± 0.5 °C with the photoperiod control of 12 hour light:12 hour dark regime. Recirculation water systems were set up in to the facility including all tanks for the filtration of the water before cycling back through the fish rearing tanks. Controlled-continuous water flow and aeration was in operation within each tank 24hrs per day. Regular checks for water quality parameters, particularly for dissolved oxygen, ammonia, nitrite and nitrate components were carried out on daily basis.

Anaesthesia was performed with care, stress was minimised as far as possible. Nets of proper mesh sizes used to capture the fish, which were then transferred into a bucket filled with water from storage tanks (28 °C) containing anaesthetic, benzocaine (ethyl-4-aminobenzoate, Sigma-Aldrich, UK) solution at the final concentration of 1:10,000 (V:V). A stock solution was first prepared by dissolving benzocaine powder at 10% (w/v) in ethanol. Whenever fish handling required including gamete collection, tagging, fin clipping; the fish were immersed into a bucket of afore-mentioned concentration of fresh benzocaine solution until the fish lost equilibrium followed by stopped opercular movement. After the required procedure was carried out, the fish were moved back to the same aquarium with water flow and aeration, then monitored until fully recovered from the effects of anaesthetic. Nets used throughout the experiments were soaked in disinfectant (Total Farm Iodophor; approved for animal health use) before and after use. Experimental broodstock fish were tagged by a TROVAN Passive Integrated Transponder (PIT) tags with a ten-digit unique code. Fish were anaesthetised prior to procedure and tagged from the lateral-abdominal side of the fish by using a special wide-type syringe (previously disinfected with 70% EtOH). Following tagging each

brood fish, females were held in separate rectangular glass aquaria to observe maturity state. The obvious signs of female readiness to spawn were swollen abdomen carrying eggs to be released with reddish genital papilla. Additionally ready to spawn Nile tilapia female, as being a part of maternal mouthbreeder genus (*Oreochromis*), showed nest building behaviour followed with eggs picking behaviour from the substrate. Female Nile tilapia can spawn at 15-20 days intervals; therefore every glass aquaria had a card attached to keep the track of spawning frequency.

2.2 Production of haploid and meiotic gynogenetic Nile tilapia

A series of experiments was carried out for the production of the haploids and meiotic gynogenetics depending on the availability of good quality eggs throughout 6 months. Each experiment, also, included a control group where ordinary fertilisation was induced. The Sarder et al. (1999) protocol was used for the inactivation of paternal DNA. Fertilised eggs were placed in an egg incubation system where they were allowed to develop normally after the period of chromosomal set manipulation applications under appropriate conditions. Observations on embryo developmental morphology and survival (24hrs, 48hrs, 72hrs, 96hrs and 120hrs AF) were used as indicators of successful application of chromosome set manipulations. The haploid group provided a unique control of UV irradiation in all experiments, suffered from so-called *haploid syndrome* thereby survived until hatching to first feeding stage maximum, compared to meiotic and bi-parental control group. Sampling was carried out in the window of hatching to prior to first feeding to ensure haploids were sampled with the meiotic gynogenetic group.

2.2.1 Collection of gametes

The whole procedure was carried out in the wet lab of the TA. First, the ovulated female was anaesthetised and placed on the bench covered with wet tissue paper. The eggs were collected by applying a slight ventral pressure (stripping) and placed in to a clean, sterile 100 mm diameter plastic petri dish. Particular attention was paid to pick the right female - ready to spawn with good quality eggs (over ripe or whitish looking batches were discarded). The eggs were washed several times with water from the egg incubation system (28 °C) to remove ovarian fluids, mucus, faeces, scales and possible blood in some cases, and held with enough water to cover all eggs. Right after egg collection, female was immediately returned to the aquarium and held under the water inflow until recovered. In the meantime, a male fish was placed into anaesthetic solution. After a few minutes, once the fish was fully under the effect of an aesthetic, the male fish was placed on wet tissue paper. Urogenital pore of the male was cleaned with wet paper tissue and gently stripped to drain any urine. Then a series of 3 to 4 glass capillary tubes (1 mm diameter, 100 µL volume Drummond Scientific Co. USA) were placed against the urogenital pore of the male while applying gentle ventral pressure. This helped avoiding activation of sperm. Any milt contaminated with urine, water or other contaminants was discarded. Once milt was collected, the male fish was returned to the original tank, placed under the fresh water supply and watched until full recovery occurred.

2.2.2 UV irradiation of sperm

Collected sperm was, first, checked for motility under a light microscope at a magnification between x10 and x25. A tiny drop of sperm was placed on a clean glass microscope slide using a micropipette tip and a drop of incubation water (28 °C) was added to activate spermatozoa. The high swimming activity of spermatozoa was the

indication of motile sperm, therefore used for the fertilisation later. Sperm was diluted down to $2.5 \times 10^7 \text{ ml}^{-1}$ concentration (Sarder et al., 1999) by using Modified Fish Ringer's Solution (MFR, pH: 8.3, stored at $+4^\circ\text{C}$ see, Ch 2.3 Appendix), to ensure proper UV irradiation throughout. A haemocytometer was used for the sperm count. Serial dilutions were made to create highly accurate reduced concentrations of sperm for the ease of counting. Although counting is the basic method of assessing sperm number with an acceptable variability due to the dilutions and counting errors, this process was reported to carry 6% error for 300 spermatozoa per observation when counting of v:v 1:500 diluted sperm was triplicated in turbot (Suquet et al., 1992). In total, 3 eppendorf tubes were labelled respectively from 1 to 3 for the serial dilution of milt. *Tube-1* was carrying all collected good quality sperm, with no dilution. *Tube-2* was the (v:v 1:10) dilution of sperm from the *tube1* (10 μL milt from *tube-1* mixed with 90 μL of MFR solution pH:8.3. Finally, *tube-3* consisted of (v:v 1:10) dilution from the *tube-2* (10 μl of diluted milt from *tube-2* mixed with 90 μL of MFR to give a final volume of 100 μL with 10x10 dilution factor). Sperm was well mixed within each step of dilution to ensure homogenous mixture and a new micropipette tip was used for each series of dilution. In total, 6-8 μL of diluted sperm from *tube-3* (dilution factor of 10x10) was placed carefully on each side of haemocytometer under a coverslip. After letting spermatozoa to be settled down for a few minutes, haemocytometer placed under the light microscope and counted in 5 large squares as indicated below (see Figure 2.1). In order to avoid any counting errors on haemocytometer, the cells that are on or touching the top and left lines of square were counted (indicated as red lines in Figure 2.1), while the ones on or touching the right or bottom lines was ignored (indicated as blue lines in Figure 2.1). The concentration of sperm was counted as follows, e.g:

Both chambers of haemocytometer were counted and average was taken:

1. Total number of spermatozoa in 5 large squares was 788 in one chamber and 650 in the other one
2. Average number of spermatozoa in 5 large squares was $= (788 + 650) / 2 = 719$
3. Average number of spermatozoa in small squares (indicated as yellow box in Figure 2.1) was: $719 / 80 = 8.9875$
4. Total concentration of sperm was : $8.9875 \times (\text{chamber volume } 4000 \times 1000) \times (\text{dilution factor } 10 \times 10) = 3.595 \times 10^9 \text{ ml}^{-1}$
5. Since optimised UV irradiation concentration of sperm was adjusted to $2.5 \times 10^7 \text{ ml}^{-1}$, the dilution was made as follows:
 $(2.5 \times 10^7 \text{ ml}^{-1} / 3.595 \times 10^9 \text{ ml}^{-1}) \times 2000 \mu\text{l} = 13.90 \mu\text{l}$ of dry sperm from *tube-1* diluted with MFR to give the total volume of 2000 μL (amount of sperm used for the fertilisation of each batches, including UV treatment groups and bi-parental controls).

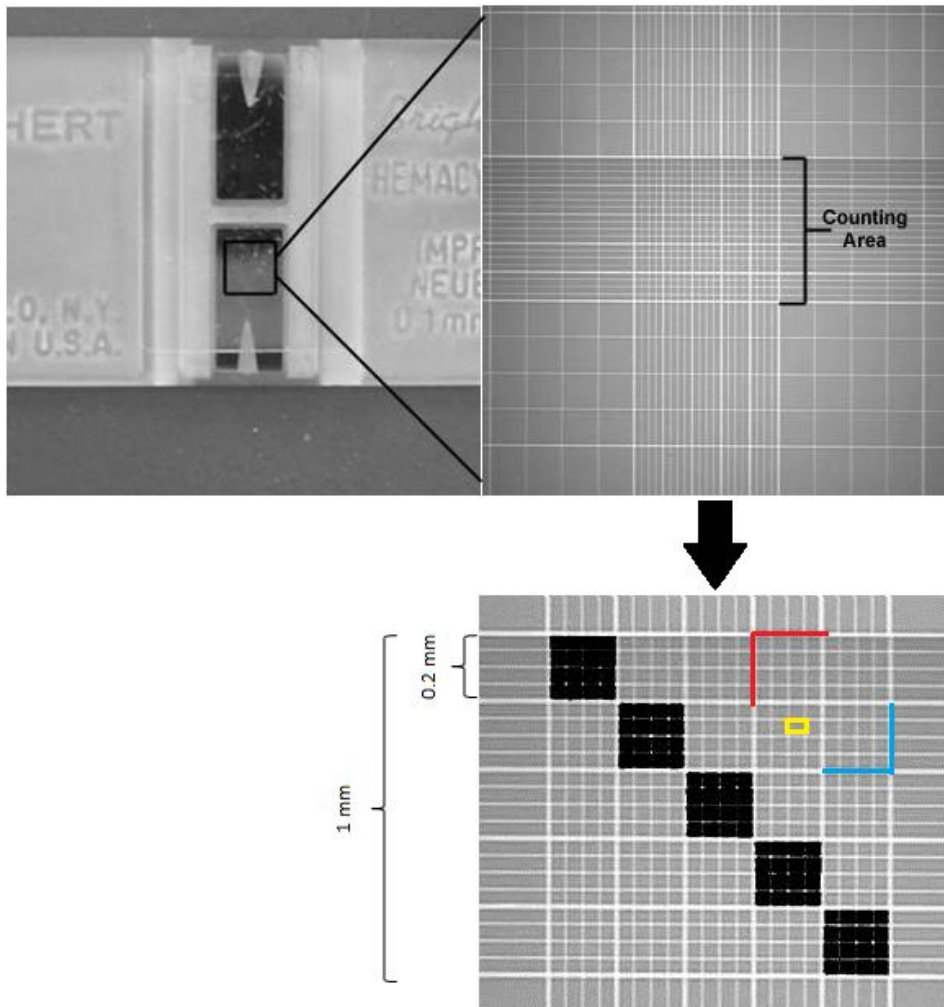


Figure 2.1: Counting the concentration of sperm using haemocytometer: Top left image shows the counting chambers with a zoom on the top right side. Bottom image represents the counted cells in 5 large squares.

The diluted aliquot was sperm placed into individually labelled plastic petri dishes (30 mm diameter) which had previously scrubbed with a scotch brite kitchen cloth to remove hydrophobicity of the petri dish, so that the sperm suspension spread evenly across the petri dish rather than accumulating in a particular area. The depth of sperm solution was approximately 1.2 mm thickness. UV irradiation was carried out in a perspex cabinet (Figure 2.3) where the petri dish carrying diluted sperm was placed on a mechanical shaker with gentle stirring and was exposed to 2 minutes of UV irradiation at the dose range of 250 – 265 $\mu\text{W cm}^2$ by using a 254 nm wavelength UV

lamp (Ultra-Violet Products, San Gabriel, California, USA). The UV lamp (Figure 2.2) was switched on at least 15 min before the onset of irradiation and UV incident dose was verified at the beginning of each experiment by using a UV radiometer (Ultra-Violet Products, San Gabriel, California, USA). The distance between UV lamp and the surface ranged 18-19 cm between experiments to achieve the desired dose rate.



Figure 2.2: UV cabinet unit used for the sperm genome irradiation.

2.2.3 Fertilisation

Fertilisation was performed after the irradiation of sperm. Good quality eggs, submerged in 28 °C incubation water, were divided into 3 batches labelled as *bi-*

parental control group (2n-Control, ordinary fertilisation), *haploid group* (n-UV Control) and *meiotic gynogenetic group* (UV+heat shock applied later to prevent the loss of polar body II) each in a separate petri dish (see Figure 2.3 for the schematic representation of the experimental design). The water was removed from the eggs, and then the eggs were fertilised *in vitro* by adding the same amount of sperm (varied between 10-14 μL between experiments) from *tube-1* topped up with MFR solution to 2000 μL to ensure the same concentration within batches. In total, 10mL of incubation water (28 °C) was added into each group to activate sperm and mixed gently. All groups were kept in incubation water until heat shock was applied to the meiotic gynogenetic group, then all groups were placed into separate-labelled egg incubators.

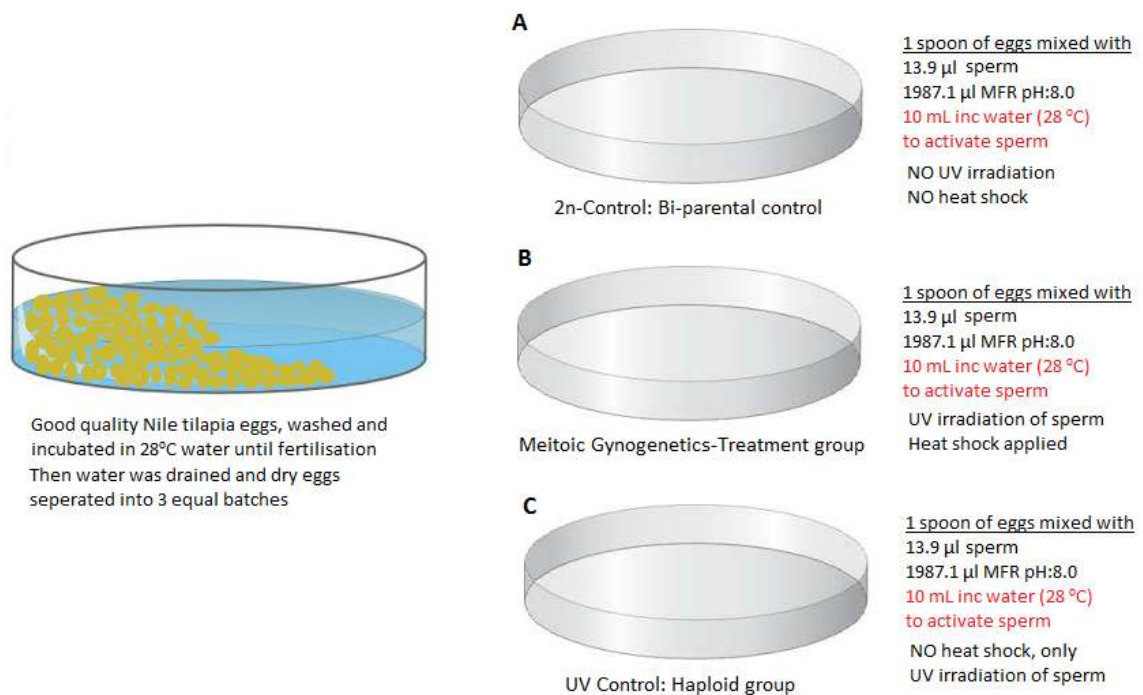


Figure 2.3: A schematic diagram for the production of control groups and meiotic gynogenetic groups.

2.2.4 Application of heat shock to fertilised eggs for meiotic gynogenetic production

The heat shock methodology followed here was as described by Sarder (1999). A temperature controlled water bath (Jencons Scientific Ltd, UK) having both cooling and

heating properties with a wide range of temperature window (-20 °C to +100 °C) was used for the heat shock application to fertilised eggs. The water bath was filled up with clean water and heated up to 41.5±0.5 °C, the required temperature, before the experiment was started. A fine mercury thermometer (1-100 °C) with 0.1 °C graduation was also used to double-check the actual temperature of the water in the water bath. After about 1 minute to allow UV inactivated sperm to fertilise eggs in the petri dish, eggs were transferred into a tea strainer and hold in a bucket of incubation water (28 °C). A heat shock was applied to developing embryos 5.0 minutes after fertilisation at 41.5 °C for the duration of 3.30 minutes. In order to ensure all eggs in the tea strainer were exposed equally to the heat shock, tea strainer was moved gently into the water bath with up-down movements. When the shock period was over, the strainer with eggs was moved immediately back to incubation water bucket (28 °C) with up-down movements and finally meiotic gynogenetic group was also placed into a separate incubator and the haploid and the bi-parental groups.

2.2.5 Incubation of eggs and sampling

A few minutes after treatments were finished eggs were washed with fresh aquarium water (28 °C) and transferred into a series of 750 ml round-bottomed plastic jars (custom-made from soft drink bottles) for egg incubation (Figure 2.4). These jars were connected to a recirculating system where warm water was fed from a 125 litre overhead tank to the jars by gravity. The water from the overhead tank first passed through a 30 W UV sterilisation unit (flow rate 20 L/min, UV dosage 62,000 $\mu\text{W cm}^{-2}$), then through 20 mm PVC pipe to the jars. They received water from the PVC pipe flow via 4 mm diameter Perspex tubing connection and the flow in the jars was controlled by small airline taps in such a way that the eggs in the jars were kept in gentle motion at all times. The system used for egg incubation was to imitate mount-breeding behaviour of

Oreochromis genus. The wastewater was discharged into the biofiltration tank (180 L capacity) via two filters filled with fibre wool positioned just above the settling tank. Shell filters helped to maintain the pH of the system and act as a surface for bacteria. The initial number of eggs was recorded and dead eggs and embryos were removed by siphoning. The embryos in each batch were checked and counted at the pigmentation stage (40-42hrs after fertilisation) and survival rate was calculated as follows:

Survival (%) = (number of embryos surviving at given development stage/total number of eggs) x 100

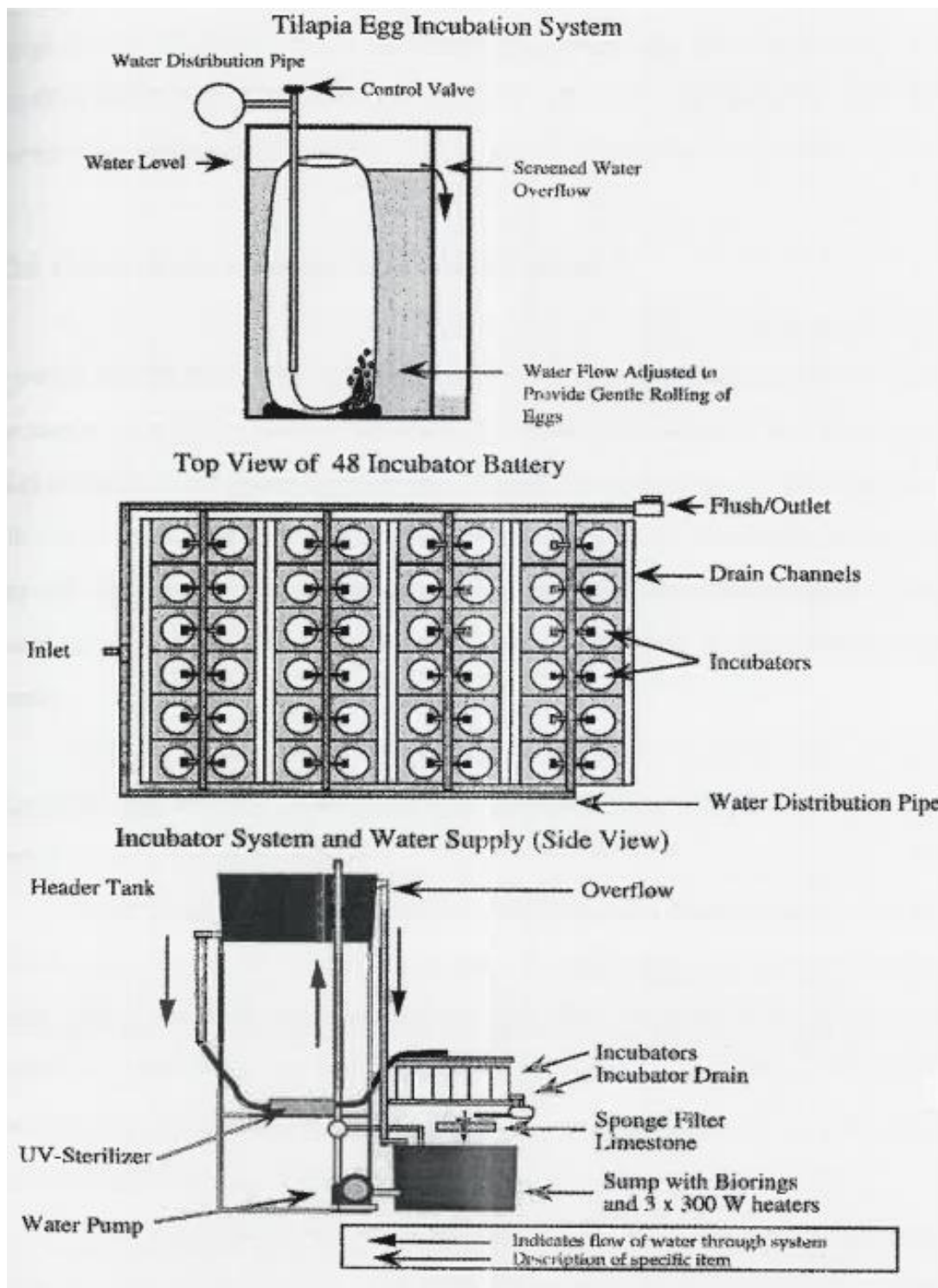


Figure 2.4 Egg incubation system set up for Nile tilapia in Tropic Aquarium at University of Stirling. (Adopted from McAndrew et al., 1995).

Fertilized eggs were kept in hatching incubators for a total period of 54 days (96hrs AF) before they were sampled. Sampling was carried out between 96-120hrs AF, until the

last stage which haploids could survive. All larvae from each batch placed into a labelled petri dish and checked under a dissecting microscope. In total, 1-2 drops of benzocaine stock solution was added into the petri dish prior to sampling of each larva into separate tubes containing absolute ethanol. Fin samples from brood fish were then collected by simply clipping 2 mm from the edge of caudal or dorsal fin without effecting swimming behaviour, using a sterile scissors, under anaesthesia. Each sample was put into microfuge tubes consists ethanol and store at +4 °C until DNA extraction. Remaining larvae from the ordinary fertilisation group were handed to the Named Animal Care Welfare Officer (NACWO) of the TA.

2.3 DNA extraction protocols

Two extraction techniques were used:

- i. REALpure DNA extraction protocol (Purification kit by REAL laboratories, Spain)
- ii. SSTNE DNA extraction protocol – universal salt extraction (Aljanabi & Martinez 1997, Taslima et al 2015)

The first method is a kit-based DNA extraction technique. However, as the protocol was not cost effective for the increasing number of samples, this method was replaced with SSTNE DNA extraction protocol, a universal, rapid salt extraction protocol, which then was used extensively throughout the research work.

Fin clips were used for the genomic DNA extraction from parents while whole larva were used for the haploid, meiotic gynogenetic and bi-parental control groups. The yolk sac was removed where possible (in most of diploid groups, including bi-parental controls and meiotic gynogenetics) however larva in the haploid group were squashed together with the yolk sac.

2.3.1 REALpure DNA extraction protocol

Initially, total genomic DNA was extracted using the REALpure kit (REAL laboratories, Spain). In total, a maximum of 24 samples were processed in the same run due to centrifuge limitation and handling. Each sample (approximately 20mg fin clip from both parents and the whole larva) was placed in an individual nucleic acid free 1.5 ml eppendorf tube containing 200 μ L or 120 μ L lysis solution respectively for parental or offspring DNA. Samples were chopped into smaller pieces using a sterile scalpel on a petri dish after removing excess ethanol on absorbent paper towel. In total, 5 μ L or 3 μ L Proteinase K (10mg/ μ L) was added respectively for parental and offspring DNA extraction into each tube and samples were overnighted at 55 °C on a rotating incubator (Techne Hybridiser, Bibby Scientific, UK) until total lysis occurred. The next day, samples were removed from the incubator and allowed to cool down to room temperature. Three μ L RNase (2 mg/ml) were added to each tube, mixed by vortexing and then samples were incubated at 37 °C for 60 minutes. Samples were brought to room temperature, 100 μ L or 60 μ L protein precipitation solutions were added respectively to parental and offspring tubes and vortex mixed. Then samples were centrifuged at 21,000 xG for 10 minutes. Precipitated proteins formed a pellet while the supernatant carrying DNA was pipetted into a new tube containing the same volume of isopropanol (250 μ L or 150 μ L for parental and offspring DNA) and mixed by 5-6 sharp (rapid and abrupt) inversions to ensure precipitated DNA in the existence of isopropanol will not be stuck to the lid of the tube. When the DNA precipitate was clearly visible (parental samples) lower speed centrifuge was used for a shorter time (>12,000 xG for 2 minutes) to pellet gDNA. The offspring higher speed centrifuge was used for a longer time (21,000 xG for 10 minutes) due to invisibility of DNA precipitate to produce a gDNA pellet. Supernatant was removed by pipetting and a quick pulse was

used to remove the last traces of excess isopropanol from each sample tube. Two 600 μ L ethanol washes were performed with freshly made 70% EtOH for the both parental and offspring samples. Each EtOH wash was left for at least 2hrs (or overnight where convenient). The tubes were centrifuged at 14,000 xG for 2 minutes. Ethanol was poured off where the gDNA pellet was visible, if not a pipette was used to remove EtOH. The same procedure was repeated once more. Then the sample tubes were first air-dried for 20 minutes by keeping them upside down on absorbent paper, and then they were placed on the hot block set at 50 °C for 10 minutes (tube lids open) to ensure each sample was completely dry. For the hydration of the gDNA 15 μ L of 5mM Tris (pH 8.5) was used for the offspring, while parental gDNA was dissolved in 50 μ L of 5 mM Tris (pH 8.5) and all samples were flick-mixed, then overnighted at +4 °C.

2.3.2 SSTNE DNA extraction protocol

The freshly prepared SSTNE buffer (Cp 2.1; Appendix) was autoclaved and stored on the lab bench at RT. Each sample (approximately 20 mg fin clip from parents or whole larva) was placed into individual nucleic acid free 1.5 ml eppendorf tubes containing 200 μ L SSTNE buffer and 20 μ L 10% SDS (sodium dodecyl sulphate-anionic detergent) was added to each tube. 5 μ L Proteinase K (10mg/ μ L) was added into each tube and samples incubated overnight at 55 °C in a rotating incubator (Techne Hybridiser, Bibby scientific, UK) until total lysis occurred. The next day, samples were removed from the incubator and placed into a hot block set to 70°C for 15 minutes to inactivate Proteinase K. Then, samples were left on the bench for a few minutes to cool down. 5 μ L RNase (2 mg/ml) were added to each tube, mixed by vortexing and then incubated at 37 °C for 60 minutes. Samples were cooled to room temperature, 158 μ L (0.7 x vol) 5M NaCl was added into each tube, mixed by high speed vortexing and left on ice for 10 minutes. Then samples were centrifuged at 21,000 xG for 10 minutes.

Precipitated proteins formed a pellet while supernatant carrying DNA was pipetted into a new tube containing the same volume of isopropanol (250 μ L) and mixed by 5-6 sharp (rapid and abrupt) inversions. When the DNA precipitate was not clearly visible, the samples were left on ice for 5 more minutes and then centrifuged at >18,000 xG for 10 minutes to produce a gDNA pellet. The supernatant was removed by pipetting when a gDNA pellet was not visible; in the case of visible gDNA pellet, the supernatant was carefully poured off and a brief spin was used to allow removal of the last traces of excess isopropanol from each sample tube by pipette. Two ethanol washes were performed with freshly made 70-75% EtOH, this time in a total volume of 1000 μ L for the both parental and offspring samples as previously described in section 2.4.1. The hydration volumes for parental and offspring samples were kept the same as explained in section 2.3.1.

2.3.3 DNA quantification and standardisation

The purity and the concentration of the extracted genomic DNA were quantified by using a NanoDrop (ND-1000) spectrophotometry (Labtech International, Uckfield, UK). These constituted stock DNA solutions. Dilutions were made from stock DNA solutions down to 50 ng/ μ L using 5mM Tris (pH:8.5), used as working solutions based on nanodrop readings.

In the spectrophotometric assays, the sample is exposed to a UV light where the absorbance read by a photo-detector. The higher the nucleic acid concentration in the sample, the more light is absorbed by the sample. There are two ratios, A₂₆₀/A₂₈₀ and A₂₆₀/A₂₃₀, commonly used to detect the purity of nucleic acids and any protein contaminations in given sample respectively. An acceptable value for the good quality DNA sample for A₂₆₀/A₂₈₀ is approximately 1.8-2.0, while expected A₂₆₀/A₂₃₀ values are commonly in the range of 2.0-2.2. Since absorbance at 260nm allows

detection of all nucleic acids in a given sample, including dsDNA, ssDNA and RNA, total absorbance of the sample can be higher than the actual amount of dsDNA, which is the template to construct ddRAD library. Therefore, accurate quantification of dsDNA was measured by fluorometric assay, Qubit® dsDNA HS Assay Kit (Invitrogen, UK) prior to the ddRAD library construction to allow dilutions to the final dsDNA concentration of 7-10 ng/μL per sample.

The molecular weight of the DNA was assessed using agarose gel electrophoresis. Briefly 1% agarose gel (70 mg agarose dissolved in 70 ml TAE buffer using microwave) was prepared and 1.4 μL EtBr (5 ug/ml) was added into gel once the temperature of the agarose decreased down to 55-60 °C under the fume hood and subsequently poured into the gel solution tray (for 24-96 samples). Combs were placed after pouring the gel depending on the number of samples and left under the fume hood to set before loading samples. Samples were prepared to load on to gel: 1 μL DNA from working solution (50 ng/μL) was mixed with 7 μL 1x DNA loading dye (Ficoll based) to fill over $\frac{3}{4}$ of the well volume (9 μL). The comb was removed carefully; the gel was placed in a buffer tank covered with 0.5x TAE buffer and samples were loaded carefully. λDNA/HindIII was used as a marker to assess the molecular weight of the samples. First, λHindIII (500 μg/ml) stock was diluted down to 100 ng/μL by dissolving 10μL stock solution of λHindIII in 40 μL TE, then incubated at 65°C for 10 minutes. In total 1.25 μL λHindIII (100 ng/μL) marker was mixed with 7 μL 1x loading dye and loaded into the first well of each raw (comb) on the gel. Genomic DNA was expected to have a high molecular weight with a single band present above 23 Kb (the biggest fragment of the λHindIII marker) without any significant DNA degradation or RNA contamination.

2.4 ddRAD library protocol

The procedure used to construct ddRAD libraries was similar to that of Peterson et al. (2012) but differed in one key matter. Early pooling of the samples was the key modification applied in the protocol used here, following the individual restriction enzyme digestion and ligation of the adapters. This modification helped to reduce the variation in the number of ddRAD reads produced at the end of sequencing run, due to reduced pipetting errors and clearly accelerated the wet lab protocol timeline compared to processing each sample singly throughout. Figure 2.5 represents the workflow comparison of the two protocols on construction of ddRAD libraries.

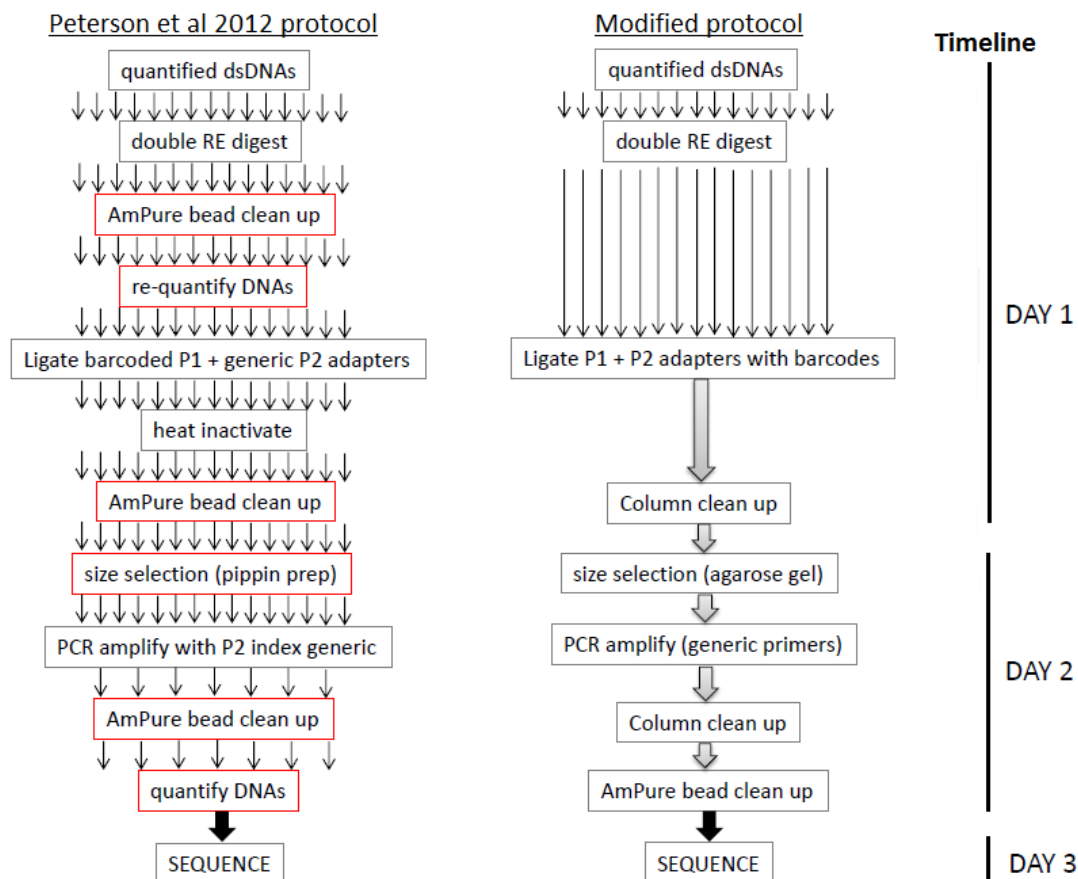


Figure 2.5: Workflow comparison of the two ddRAD library protocols. Peterson et al. (2012) is the original ddRAD protocol, the protocol on the right is the modified protocol used to construct ddRAD libraries throughout the thesis.

In total, 7 ddRAD libraries were constructed in European seabass, Atlantic salmon and Nile tilapia. Each ddRAD library varied from one another in terms of number of samples used, although the protocols used within experiments were almost identical. The wet lab protocol of ddRAD library had to be modified in some cases where genomic DNA available was a limiting factor or depending on the yield of library PCR cycles varied among libraries (minimum DNA concentration used was 12 ng DNA per sample in European seabass androgenetic samples-not included in the thesis).

Throughout ddRAD library construction protocol, care was given to avoid any cross-contamination of the consumables, particularly adaptor mixes. All consumables that needed to keep cool were kept on ice in a polystyrene rack. All master mixes including restriction digestion, adaptor ligation and PCR were prepared in an extra 20% volume.

2.4.1 Restriction digestion

Restriction enzyme double digest reaction involved *SbfI* as a rare cutter with an 8 bp recognition site (CCTGCA|GG motif) and *SphI* as a common cutter (GCATG|C motif). Both enzymes were supplied by New England Biolabs, NEB and neither of them was methylation sensitive, nor they had star activity as they were High Fidelity (HF®) versions of the native enzymes. An *in silico* exploration was carried out on available fish genomes before deciding on the combination of enzymes. This investigation revealed that *SbfI* & *SphI* combination produces 2,500 to 4,000 fragments (an appropriate number of ddRAD tags) in the desired size range of 300-600 bp (optimal for Illumina sequencing).

A day before starting to construct ddRAD libraries, quantification of the dsDNA was carried out; final dilutions were made x5 more in volume in case of extra DNA required in the same concentration as above explained. On the first day of ddRAD library construction, mastermix (10X Cut Smart Buffer, NEB) was thawed in RT and REs were

placed on ice in a polystyrene rack to avoid any contamination from the ice. Fresh ddH₂O was placed into a nuclease-free Eppendorf tube and placed on ice. All master mixes and consumables were flick-mixed and centrifuged briefly to ensure homogeneity mixture. Individual reactions were set up in a nuclease free 96 well plate pipetting 3 µL of 7 ng/µL standardised DNA into the plate by using a multichannel pipette. Double-digest restriction master mix was prepared in a nuclease free Eppendorf tube containing 10U of each enzyme per microgram of genomic DNA in 1x cut Smart BufferTM (NEB). Restriction digestion reaction was set up in a way that an equal volume of genomic DNA was mixed with an equal volume of mastermix – 3 µL of each, to give total reaction volume of 6 µL per individual sample. Master mix was added into individual reactions on the 96-well plate and pipetted up and down twice to ensure homogeneity mix which would produce more even RE digestion. The plate was sealed, mixed and centrifuged twice, then placed into a thermal cycler at 37°C and incubated for 40 minutes.

2.4.2 Ligation adaptors

The P1 and P2 adaptors were designed in a way that P1 was compatible with the *Sbf*I overhang and P2 was compatible with the *Sph*I overhang. 5 bp to 7 bp barcodes were included in the adaptor right after (3' prime to) the Illumina sequencing primer. This design ensured the first base read by Illumina would be the barcodes which was followed by the remainder of the RE site (still a part of genomic DNA) and the genomic DNA of interest (see Figure 2.6).



Figure 2.6: Design of the adaptors used for the construction of ddRAD library and the structure of the ddRAD library fragment formed by initial ligation of a *SbfI* P1 adapter (5bp barcode is blue colour coded) and a *SphI* P2 adapter (5bp barcode is blue colour coded).

While genomic DNAs were still digesting, ready to use, numbered adapter/barcode mixes were removed from the freezer and placed into a fridge where gentle thawing took place (avoiding causing a thermal shock by letting them sit at room temperature). Each adaptor mix contained a unique set of P1+P2 combinations which later was used to de-multiplex raw reads. The numbers of the adaptors corresponded with the numbers of the individual samples (e.g.: row A1 to A10 of the adaptors was mixed with row A1 to A10 in 96-well plate). Once restriction digestion was completed the 96-well plate carrying individual reactions were removed from the thermal cycler and placed on to the bench where they were held to cool down to room temperature. T4 ligase (2 M ceU/mL, NEB) and rATP (100 mmol/L, Promega) to be used in the adapter ligation mastermix were removed from freezer and placed on ice. Gently thawed adapter mixes were briefly centrifuged to make sure all solutions were in the bottom of the tubes for ease of using the multipipette. A PCR thermal cycler with the heated-lid set at 95 °C

while the block temperature was 22°C, was kept on throughout the ddRAD library construction to remove plastic seal easily from the 96-well plate carrying individual reactions. In total, 3 µL of adapter/barcode mix (*SbfI:SphI* 1:10) was added to each RE digested genomic sample by pipetting up and down to ensure even mixture. Adapters and RE digested samples were incubated for 10 minutes to allow initial annealing of sticky ends while adapter master mix was prepared [0.15 µL 100 mmol/L rATP (Promega), 0.25 µL 10× CutSmart™ Buffer (NEB), 0.03 µL T4 ligase (NEB, 2 M ceU/mL)] added. Reaction volumes were made up to 12 µL with nuclease-free water for each sample (3 µL per individual reaction). The 96-well plate was sealed, mixed, centrifuged twice to ensure homogenous mixing and incubated at 22 °C for over 3hrs (this gave desired interval to set up agarose gel for size selection for the next day).

2.4.3 First purification step

MinElute spin column PCR purification kit (Qiagen) was used for the purification of digested and adaptor - ligated genomic DNA fragments. Following 3hrs of adapter ligation, samples were briefly centrifuged; the heated lid was used to remove the sticky lid gently from the 96-well plate. A 3x volume of PB buffer (36 µL) was added to each individual sample and all samples were pooled into 7mL clear tubes where the pH of the solution was observed (light yellow colour was required for the optimal pH [$<$ than 7.5 is required for the efficient adsorption of the DNA to the membrane] of the solution). Depending on the colour of the solution 2-3 µL of 3 M sodium acetate (NaAc) (pH 5.2) was added to the pooled samples (colour change observed from dark orange to light yellow), mixed well and processed through a single Qiagen column using 550 µL sequential aliquots of PB buffer / DNA mix. The DNA was eluted in 2x of 65 µL of warmed (65 °C) Qiagen supplied EB buffer, to obtain 125 µL volume in

total which was stored on ice until the next day to run on agarose gel for the size selection.

2.4.4 Size selection

The gel (1.1% - 0.42 g agarose dissolved in 38 mL 0.5xTAE buffer with no EtBr, c.6.5 mm thick) for the size selection was poured on the first day of the ddRAD library construction, during the adapter ligation step and stored in a fridge, submerged in 0.5x TAE buffer. A comb with a single well to hold 200 μ L of template (made by taping across several teeth to form one large well with autoclave tape) and 2 flanking wells to hold 10 μ L of 320-590 bp markers on either side of the large well were formed the gel. Freshly made 0.5x TAE buffer, used for the gel, was also chilled in the fridge as well as the electrophoresis apparatus.

On the 2nd day of the ddRAD library construction, first, the gel electrophoresis system was set up on ice with chilled buffer. The idea of applying a chilled run was to minimise the diffusion of the smaller fragments in the gel and get more precise sample fractionation. The gel tray was added to the electrophoresis system, just submerged with chilled fresh buffer and a test-run was applied for 10 minutes to ensure the electrical contacts were sound. To specify the target range for selection, 2x marker reactions were prepared containing 2 μ L Marker I (590bp), 2 μ L Marker II (320bp), 1.8 μ L 6x LD and 6.2 μ L of EB buffer, and loaded into the gel to see that the wells were sound before loading the template. Then, 125 μ L DNA template was mixed with 25 μ L 6X DNA loading dye to achieve 1x final dye concentration on the gel and loaded in two batches of 75 μ L followed by a waiting for a few minutes for the template to equilibrate in the large well. Then electrophoresis was started with lower voltages (20-40-60-80 V/cm) for a few minutes and increased up to 105 V/cm for the remainder of the run. The gel was run on ice until the dye was 3.3 cm away from the origin, which took

around 1hr of electrophoresis. Once the electrophoresis was stopped, the gel was placed on a clean glass to cut out the fragments of interest, starting from the edge of the big well where the DNA template was loaded until 2-3 cm ahead of the dye. This was stored in the fridge (Figure 2.7C) until the remainder of the gel was stained in EtBr solution (2 μ L EtBR (stock concentration of 5 μ g/ml) added into 100ml dH₂O) to visualise the position of the markers to detect size of interest (Figure 2.7A). The stained gel was viewed on safeview under UV / blue light and the position of the marker bands were notched to mark the size range to cut, represented as orange rectangles at the both side of the library (Figure 2.7D). The stained gel was washed with dH₂O to remove any EtBr residues, moved back to the lab bench and the cut-out lane carrying the fragments of interest was placed on the gel. The notches made under UV light made it obvious to cut the actual size range under the fluorescent light on the bench. A sterile razor blade was used to cut off the piece of interest and the remaining gel was re-stained in EtBr then imaged for the record of restriction (Figure 2.7D).

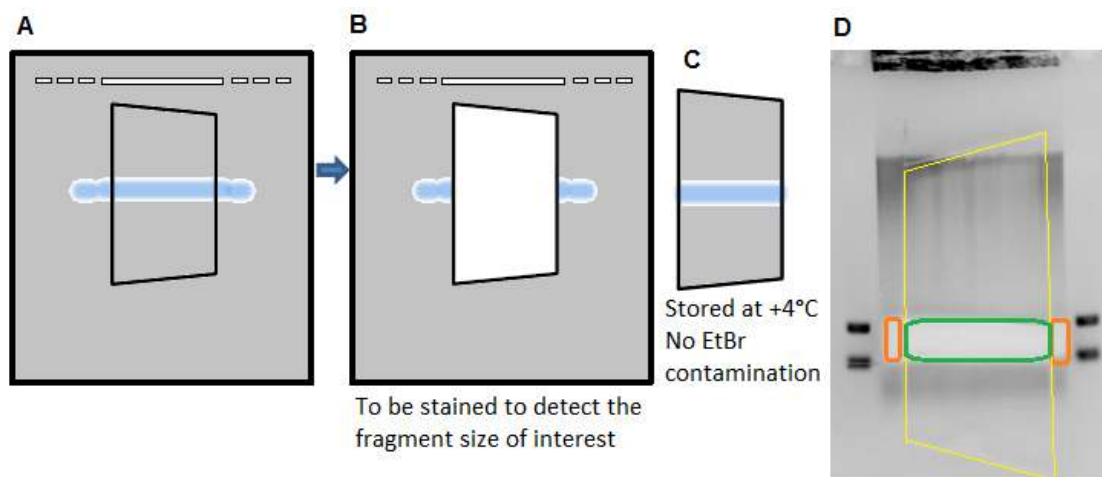


Figure 2.7: A schematic diagram of size selection on the agarose gel. A represents the whole gel at the end of electrophoresis run, B shows the cut-out lane, ensures dye stayed in the middle, C represents the cut-out lane stored in the fridge without contaminating library with EtBr until remaining gel is stained in EtBr solution and visualised under safeview UV blue light, D represents the whole gel image after cutting out fragment of interest and after stained in EtBr solution for imaging. The orange rectangle represents the notched parts to identify size range on the gel and the green rectangle represents the fragment size of interest, later used to extract template from the gel, while the yellow lines depict cut line of the whole gel corresponding to the image C on the left.

2.4.5 Second purification step

The gel slice was weighed then sliced evenly and split between 2-3 eppendorf tubes (Vol 1.5 mL) (see Figure 2.8). A MinElute spin column gel purification kit (Qiagen) was used for the purification of library from the agarose gel. For that, 3x volume of QG buffer was added into each tube depending on the weight of the gel slice (e.g: 0.28 g gel was mixed with 0.84 mL of QG buffer) and the samples were placed in a rotator derive STR4 (Stuart Scientific, UK) and allowed to dissolve at room temperature with agitation for 10-15 minutes. The tubes were briefly centrifuged and 1x volume isopropanol (e.g: 0.28 mL isopropanol) added, which was mixed and then the samples were processed through a single Qiagen column using 550 μ L sequential aliquots of QG buffer and DNA mix. The colour of the mixture was yellow, therefore did not require adjustment of the pH. The DNA was eluted in 2x 35 μ L of warmed (65 $^{\circ}$ C) Qiagen supplied EB buffer, to obtain 65 μ L volume in total.

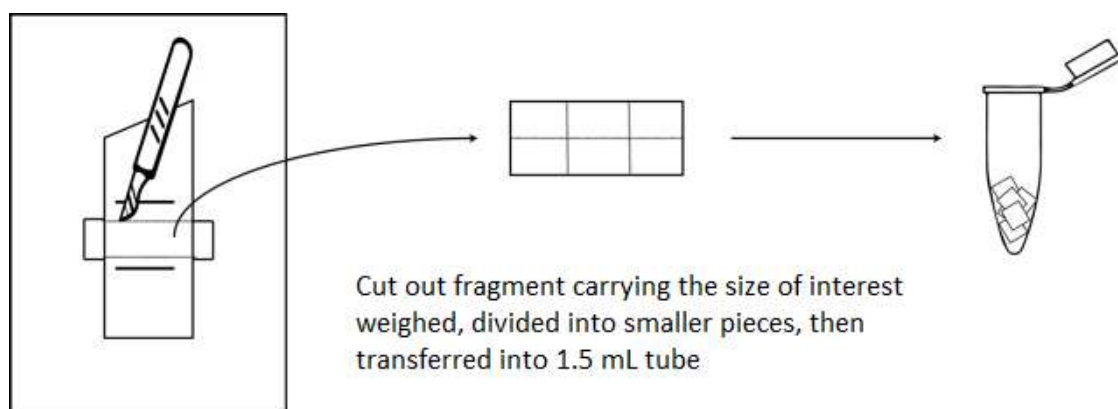


Figure 2.8: A diagram represents the initial process of getting library with the size of interest from the agarose gel for the later purification process.

2.4.6 *Enrichment of library*

Two separate PCR test reactions were set up for the optimisation of the bulk PCR reaction. The first PCR was a standard 1 μ L DNA template per 25 μ L reaction volume plus a non-template control (NTC) with both run for 16 cycles. The second PCR, run for 13 cycles, involved a standard 1 μ L DNA template per 25 μ L reaction volume and a double template (2 μ L per 25 μ L). Both the first and the second tests were performed in half reaction volume (12.5 μ L) in 0.2 mL thin-walled PCR tubes involved 0.2 μ L P1+P2 generic primer mix (10 μ M), 6.25 μ L 2x NEB Q5 HS mix, 4.05 μ L ddH₂O and remaining 2 μ L was adjusted with the template DNA and topped up to 12.5 μ L ddH₂O. Tubes were mixed and centrifuged briefly to ensure homogenous mixing. Amplification was carried out as follows: 30 sec at 98°C, 13x and 16X [10 sec at 98°C, 30 sec at 65°C, 30 sec at 72°C], 5 min at 72°C, hold at 22°C. Both thermal cyclers (one was set up for 13x cycles and the other for 16x cycles) had the heated lid set up to 98 °C and hold/pause the unsure rapid heating to desired temperature. In total 5 μ L of each amplicon of the test PCRs (mixed with 3 μ L 3x DNA loading dye) were loaded on a 1.5% agarose gel with a 100 bp DNA ladder, Gene Ruler (Figure 2.9).

The number of PCR cycles used to enrich library at between 11-14 cycles. This decision was to ensure consistency for the all ddRAD libraries constructed through the research. Too many PCR cycles (>18 cycles) are likely to introduce PCR-based errors, and may also increase the ratio of GC-rich fragments, while too few cycles (<10 cycles) may increase the ratio of AT-rich fragments.

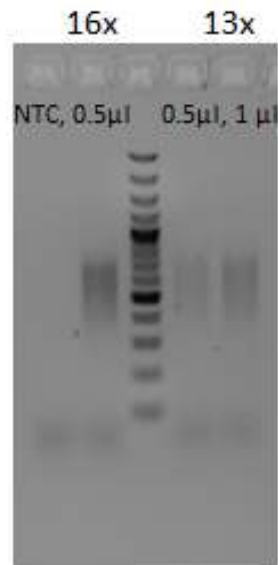


Figure 2.9: Test PCR to optimise enrichment of the library. Wells on the left side of the marker carry the amplicons of 16x PCR cycles while the wells on the right side are the amplicons of 13x PCR cycles. Well-1 shows the NTC, Well-2 has 0.5 μ L template (16x), Well-3 is 100bp Gene Ruler, Well-4 is 0.5 μ L template (13x) and Well-5 is the double template (13x).

The decision on the number of cycles was made based on the test PCR for the bulk amplification of the template (e.g. based on above example 13x PCR cycles with 1.5 μ L template for the half reaction was used, which means the intensity of the amplicon was expected to be the average of wells 4 and 5 (Figure 2.9). The amplicon of the double template following 13x PCR cycles well 5 was very strong (ideal template should be faintly visible) while the standard template following 13x PCR cycles used well 4 was rather weak for the final amplification of the library. Having selected template volume and PCR cycles that gave adequate amplifications, a large scale master mix was prepared to produce sufficient quantity of the library for the sequencing (desired

volume of 300-400 μL), and split in to as many individual PCR reactions as possible to spread any bias arising from individual reactions. Master mix (involved all the solutions and the template) was added into individual wells of the 96 well plate, run in half reaction volumes (12.5 μL). The plate was sealed, mixed, centrifuged briefly and placed into a thermal cycler. When cycling was completed, all aliquots were combined into a single tube and 5 μL of the bulk product (optional, can be check after purification, too) was checked on agarose gel.

2.4.7 Amplicon clean up (third purification)

Following bulk amplification of the library, a column-based purification was performed mainly to remove all the master mixes, enzymes, salts and dNTPs from the library but more importantly to reduce the volume of the library (32 individual reactions each in 12.5 μL volume ended up 400 μL library) for the later magnetic bead clean up (expensive consumable). A MinElute spin column PCR purification kit (Qiagen) was used for the purification of bulked library as explained in 2.4.3 and eluted in 65 μL EB buffer.

2.4.8 AMPure magnetic beads (final purification)

The paramagnetic bead approach has recently become the choice for researchers working on NGS library construction due to the reputation for providing a cleaner product. Paramagnetic means the beads are magnetic only in the existence of a magnetic field. This feature prevents them from clumping in the solution. Each bead has a polystyrene core surrounded by an extra layer of magnetite which is coated with a carboxylate-modified polymer surface. Such coating constitutes the binding surface, reversibly, for the DNA in the presence of polyethylene glycol (PEG) and salt (20%

PEG, 2.5M NaCl mix, comes with the solution). PEG stimulates negatively-charged DNA (due to phosphate groups) to bind with the carboxyl groups of the bead surface.

The size of the fragments eluted from the beads (or binding in the first place) is dependent on the concentration of PEG and salt in the reaction mix and this, in turn, is determined by the mix of bead:DNA which is very crucial. Since the ratio determines the length of fragments to be bond or left in the solution mix (see Figure 2.10) a lower ratio of bead:DNA leads to higher molecular weight DNA fragments binding to the beads, so larger fragments will be eluted at the end. This selective mechanism is due to larger fragments possessing a larger total charge per molecule, so promoting the beads to bind to DNA molecules, with larger charge rather that of smaller molecules therefore clearing behind smaller fragments from the reaction. The binding capacity of the beads is very high, e.g 1 μ L AMPure XP binds $>3\mu$ g DNA. Although these beads were mainly designed for purification of PCR amplified colonies initially (DeAngelis et al., 1995), nowadays they are mostly used for NGS library purification due to the added benefits of increasing scale and the reproducibility of the library as well as reducing the input DNA requirement (Fisher et al., 2011).

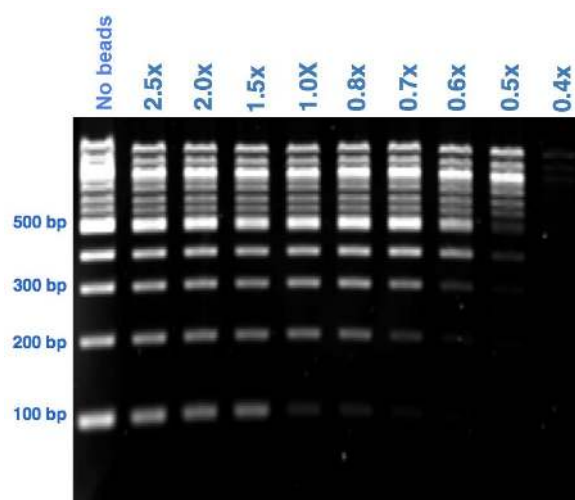


Figure 2.10: The change in the purification efficiency of paramagnetic beads as the ratio of bead:DNA decreases. Adopted from Broad Institute boot camp (URL: <https://www.broadinstitute.org/files/shared/illuminavids/SamplePrepSlides.pdf>, slide no 52).

AMPure magnetic beads (Beckman Coulter, UK) were used for the final purification of the ddRAD library (see Figure 2.11 for the schematic diagram of the clean-up procedure with magnetic beads). An equal volume (65 μ L) of paramagnetic beads removed from the fridge, equilibrated to room temperature and mixed well, was added straight into the purified library tube (65 μ L). Meanwhile, a heat block was set up to 60°C to warm 50 μ L of Qiagen supplied EB buffer. The beads and the library were mixed by pipetting gently to ensure that the solution stayed at the bottom of the tube. The tube was incubated at room temperature for 5 minutes before being placed into a magnetic stand lids open and left 3-4 minutes until the beads (brown colour) had migrated to the side of the tube where the magnet was located. From this point until eluting the final cleaned-up library, the tube stayed undisturbed in the magnetic stand and very careful pipetting technique was used throughout the protocol). Once all beads had migrated and a clear reaction mix was observed, the supernatant carrying DNA fragments smaller than 200bp was removed by careful pipetting. Fresh 73% EtOH (1 mL volume) was prepared simply by diluting 730 μ L 100% EtOH + 270 μ L nuclease free water, and two sequential EtOH washes each with 30-60 seconds incubation were performed. Then tube was carefully removed from the magnetic stand and placed into a 60 °C heat block to dry out the beads for 2-3 minutes. Then the beads were gently re-suspended in 20 μ L warm EB buffer, mixed by gentle pipetting and returned to the heat block for 2-3 minutes. Then the tube, lid open, was placed into the magnetic stand for 3-4 minutes until all the beads fully migrated to the side of the tube. All of the supernatant, carrying the library, was carefully pipetted into a sterile tube and labelled as ddRAD library.

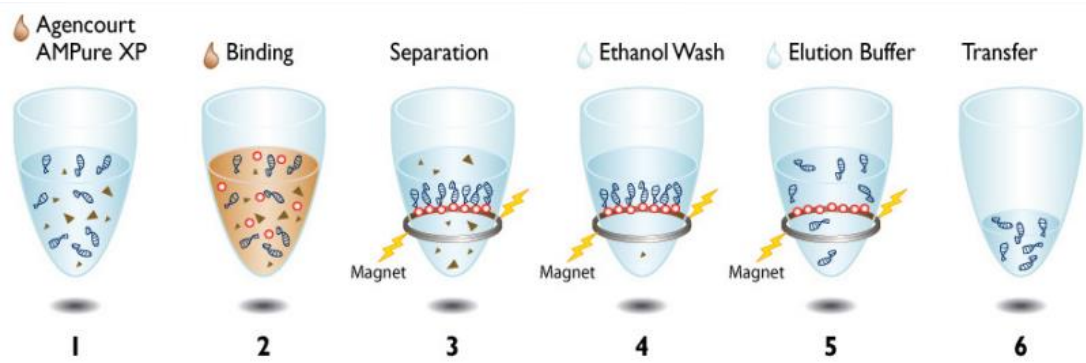


Figure 2.11: AMPure paramagnetic beads clean-up procedure, copied from Beckman’s website (URL: <https://www.broadinstitute.org/files/shared/illumina/vids/SamplePrepSlides.pdf>, slide no 51).

2.5 Data Analysis

This section provides an evaluation of initial sequence quality for high throughput data analysis which substantially affects the reliability of the downstream analysis followed with building loci and calling genotypes from the short reads produced from the high-throughput sequencer.

2.5.1 Quality Control of raw data

The overall quality of the sequencing run was initially assessed using metrics generated by the sequencer (MiSeq Control Software (MCS), Miseq, Illumina). This however, mostly focused on identifying problems which were generated by the sequencer itself. Therefore an independent software FastQC v.0.11.3 (Andrews, 2010), designed to spot any problems that can originate either in the sequencer or in the starting library material, was used to generate a comprehensive QC report. FastQC uses raw data that comes straight from the high-throughput sequencer. In total, 12 analysis modules are performed covering basic statistics to potential contaminants within the library. One important aspect that needs to be taken into account is that although analysis reports appear to give a clear pass/fail result, such evaluations need to be taken in the content of

what is expected from the library itself. As FastQC point of view, a *normal* data should be random and diverse while in the ddRAD libraries that were produced in the present thesis are expected to be biased at the beginning due to limited adaptor choices. This automatically flags up in two modules “per base sequence content” and “Kmer content” (see Figure 2.12 and therein explanation). Therefore summary evaluations were treated as pointers to focus attention mostly on per base sequence quality and what may not look random and diverse within each library for the quality control validation. Sections from now on to 2.6.1.12 were mainly adapted from the website of the software FastQC (Andrews, 2010) as well as experiences gained throughout the project.

2.5.1.1 Basic statistics module

This module generates simple statistics on file name (original file name analysed), file type (whether file has actual base calls or colour spaced data which in this case needs to be converted to base calls), encoding system used (ASCII encoding of quality scores), count of total sequences, sequences flagged as poor quality, sequence length providing the shortest and longest sequences of data and overall GC content of the library. It is not usual for this module to flag up a failure.

2.5.1.2 Per base sequence quality module

This module is the most important outcome of the entire QC report, providing an overview of the range of quality values across all bases at each position in the raw sequence data file. All of the individual bases per reads are plotted to X-axis while quality scores (Phred) are plotted on the Y-axis of the graph. As the score gets higher better base calls are indicated. The colour codes used at the background divides data into three segments; very good quality calls (green; Phred>30), calls of reasonable quality (orange: Phred 20-30) and calls of poor quality (red: Phred<20). It is expected to

see base calls falling into a level where a warning or an error is triggered as the quality of the calls on most high-throughput sequencing platforms the chemistry will degrade as the sequencer run progresses.

Phred scores are used for the assessments of sequence quality this was originally developed for the accurate DNA sequencing for the Human Genome Project (HGP) (Ewing et al., 1998). These scores are logarithmically linked to error possibilities. Table 2.1 demonstrates the base calling accuracy with increasing quality scores.

Table 2.1: Phred score quality score interpretation. Material adopted from Wikipedia.

Phred quality score	Probability of incorrect base call	Base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99,9%
40	1 in 10,000	99,99%
50	1 in 100,000	99.999%

2.5.1.3 Per sequence quality scores module

This module visualises the distribution of mean score across all the sequences to see if universally high quality values are tightly distributed. If a subset of sequence(s) falls into low quality values this can indicate a systematic error mainly due to the run, such as a part of run had a problem. A common case to raise a warning in this module is if the mean quality falls below 0.2% error rate.

2.5.1.4 Per base sequence content module

This module plots the proportion of each base position throughout the sequencing run. It is expected to have a little or no difference; therefore the lines representing each nucleotide should run almost parallel to one another regardless of the position. Although the relative amount of each base should be a reflection of the overall amount of these bases in the genome of interest, they should not represent a highly imbalanced representation in any case. A warning is raised if the difference between A and T or G and C is greater than 10% in any position. This is a normal case for all the libraries constructed within the scope of the present thesis where restriction enzyme recognition site (still part of genomic DNA) is stacked together. Owing to select the fragments carrying only RE site such bias is expected from the 5th-7th bases to the end of enzyme recognition site for the libraries constructed (see Figure 2.12).

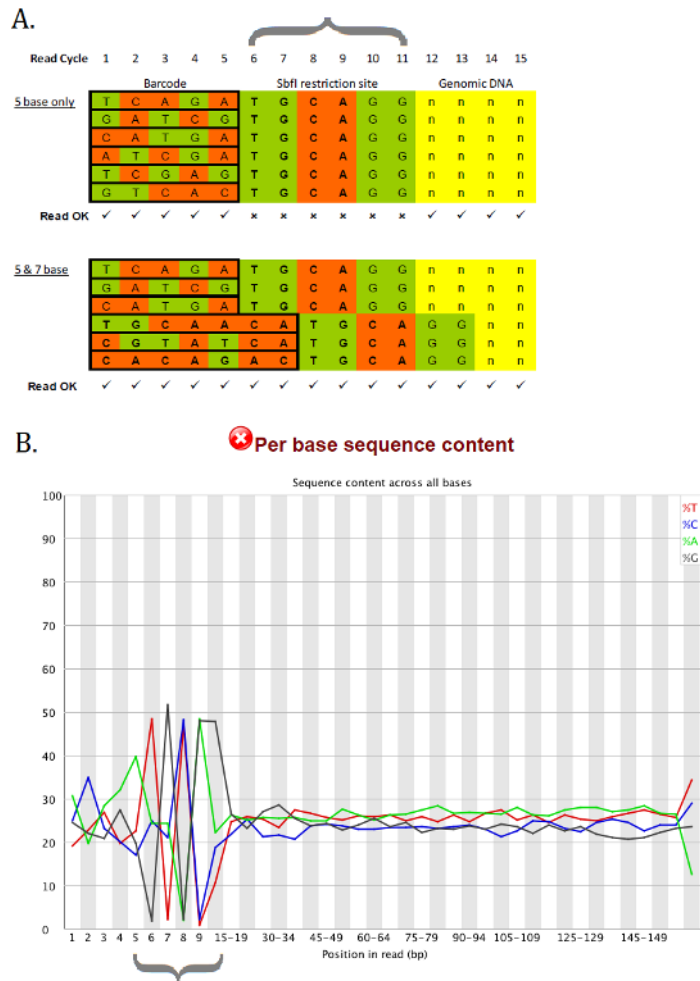


Figure 2.12: A schematic diagram to show expected fluctuations at the beginning of the sequence reads. Above figure (A) represents the distribution of the first 15 bases in a frameshift format that ensures balanced green/red laser detection for successful sequencing run on Illumina while below figure (B) shows the warning at the end of FastQC -per base sequence content module as a result of natural composition bias in the fragments.

2.5.1.5 Per sequence GC content module

This module plots the overall GC content of the library per sequence and compares it to a theoretical normal distribution model. As FastQC point of view, a central peak corresponds to GC content is expected in a diverse and random library so called *normal*. In the case of an unusual shaped distribution being observed up to 15% a warning is raised. Any contaminants (biological or chemical based) can be recognised

by the shape of the peaks which are triggered by the different molarity of bacterial genomes. Sharp-absurd peaks on a smooth distribution are normally the results of specific contaminant such as adaptor dimers while broader peaks are more likely to be a representation of a contamination from a different species.

2.5.1.6 Per sequence N content module

This module plots the frequency of failure, when the sequencer was unable to make a sufficient base call. Although it can be usual to see a few N bases appearing in longer sequences or at the end of the sequences, it is not expected to observe Ns in the short reads such as that of Illumina platform (considering Phred>30 scores are commonly in use ensuring a probability of 1 nucleotide incorrect base call in every 1000 bases, see Table 2.1). This module flags up a warning if any position represents an N content of >5%.

2.5.1.7 Sequence length distribution module

This module generates a graph showing whether the sequence length distribution is achieved uniformly or not. Considering it is entirely normal to have different read lengths for most high-throughput platforms the warning that will be raised from having non-uniform lengths can simply be ignored. All ddRAD libraries generated in the thesis involved Illumina V2 chemistry with 150bp paired end sequencing. However based on the experiences gained, all reads were sequenced up to the upper limits of the chemistry (which was 161bp for both reads).

2.5.1.8 Duplicated sequences module

This module counts the level of duplication for every read within the library and plots the relative number of sequences with a different level of duplication. The idea here is

to observe how unique the sequences within the library are. As FastQC point of view vast majority of the sequences should occur only once in a *normal* library (random and diverse). While a low degree of duplication is more likely to indicate a very high coverage of the genome of interest, a high degree of duplication may signify an enrichment bias which might either be propagated by PCR amplification during library construction or might have a biological explanation as observed in the duplicated genomes (natural stacks where different copies of exactly same sequence are randomly selected due to high existence of such regions in the genome of interest). Therefore warning arises in this module can be ignored to some extent if there is a biological explanation to the case (assuming as low PCR cycles as possible, used to avoid any technical bias).

2.5.1.9 Overrepresented sequences module

This module creates a list of overrepresented sequences which is defined to occur more than 0.1% of the total reads with their counts, percentages as well as possible source for the high representation. In order to identify common contaminants this module also involves all the primers and the adaptors of various sequencing platforms in default settings. This can also be customised with an expected source of contaminant.

2.5.1.10 Adapter content module

This module plots a generic analysis of all the adapters in the library to find out those that have even coverage in the sequence data. This plot is a cumulative percentage count of the proportion of the library. Therefore an increase is more likely to be observed through the end of the read as the read length continues.

2.5.1.11 Kmer content module

This module plots a graph to observe any unusual enrichments of the sequence in Kmer format. For the present thesis point of view, this module is treated as an extension of overrepresented sequences search which allows finding partial sequences that are overrepresented but some might not appear at a fixed position for each read. Kmers are a short form of DNA sequence (k is the DNA “words” of length) usually divisible by four. Distribution of kmers in DNA provides an interesting perspective on the complexity of the genome of interest particularly in whole genome sequencing projects.

2.5.1.12 Per tile sequence quality module

This module will only appear in the final quality report in the case of using an Illumina library which retains its original sequence identifier. These sequence identifiers are a series of systematic information gathered per read including unique name for the sequencer, run ID, flow cell ID, flow lane and tile number in the each flow lane, X and Y coordinates of flow lane, index number and sequence identifier number (which can either be 1 or 2 in which 2 refers to paired end read) followed with sequence and the quality scores (phred). Every small square represents a tile from the flow cell across all bases called to observe any particular quality drop associated with only a part of the flow cell. Any colour change from average quality, represented as blue towards hot colours such as orange, yellow and red refers to quality drop in specific tile. Therefore a good sequencing run should produce a graph of the shades of blue. A warning is generally arisen if flow cell is overloaded. Mildly effected small number of tiles can be ignored but in the case of observing larger effects which can show high deviation in score for several cycles, re-run can be choice to ensure high quality sequencing.

2.5.2 *Stacks Pipeline for building loci*

Stacks (Catchen 2011) is a bioinformatic pipeline that has been designed to work with any restriction enzyme based data including RADseq, ddRADseq or 2b-RADseq. This pipeline performs the best with Illumina platforms where short reads are produced by high-throughput sequencers. Stacks assembles massive numbers of short read sequences throughout multiple samples. Bringing such short reads together is defined as making an assembly, which is achieved in two ways: either *de novo* (no genome assembly is used or available) or reference-based (by aligning reads to a reference genome assembly, including gapped alignments). The *de novo* pipeline (**denovo_map.pl**) compares all the sequenced reads and build stacks of exactly matching tags (by applying *ustacks*, *cstacks*, *sstacks* respectively, See Figure 2.13). Pairwise comparisons are made within all stacks in such a way that each stack must differ from another by at least one base (default parameter). Then each locus is examined one nucleotide divergent position at a time and this process is repeated for each individual. The reference-based (**ref_map.pl**) pipeline on the other hand takes reference aligned input data and then generates loci and makes SNP calls by applying each stacks components (*pstacks*, *cstacks*, *sstacks* respectively, See Figure 2.13). All parameters in stacks pipeline can be customised to fit the nature of the experiment. A maximum likelihood statistical model incorporated in Stacks provides a sufficient pipeline to differentiate sequence variations and polymorphisms from sequencing errors. This pipeline can be used both for families (genetic crosses) or population samples. In the scope of this thesis, all samples were family based. The Stacks pipeline is run in two consecutive stages as below:

2.5.2.1 Cleaning and de-multiplexing data with process_radtags module

This module is the essential step for each next-generation sequencing data analysis: removing low quality sequences and separating reads from different samples which were individually barcoded within the library. This module uses the outcome of the high-throughput sequencer and first checks the barcodes and the restriction enzyme cut-sites are undamaged (correcting minor errors as default). Then the module runs through the short sequence and checks the average quality using Phred (>33) scores. If the quality score falls below Phred10, probability of incorrect base call is 1 in every 10 bases (see Table 2.1, 2.5.1.2 explaining Phred score encoding), the read is discarded to ensure retained reads to be high quality for SNP calls later. This module outputs the files according to sample names with retained reads and provides a summary table. Perl shell scripts used are available on request.

2.5.2.2 Building loci and calling SNPs (ustacks/pstacks, cstacks, sstacks)

The first task of this module starts with creating a MySQL database to store and visualise the results. In order not to have a permission issues with MySQL database the first line of Perl scripts includes ('_radtags') as a suffix (e.g: salmo_radtags). Then, based on the assembly method, the main stacks pipeline is run. The denovo_map.pl first runs ustacks on a set of samples, building loci and calling SNPs in each reads. Then cstacks is run to create a catalogue from the entire loci that have been observed in parents and sstacks is run subsequently to match all genotypes of progeny against to the catalogue of parental genotypes (Figure 2.13). There are some considerations that significantly affect the outcome of the pipeline, including: -m, -M and -n parameters. As a rule of thumb, these parameters are set to -m:10, -M:2 and -n:1. (see tutorial on how do the major stacks parameters control the *de novo* formation of stacks and loci at

http://catchenlab.life.illinois.edu/stacks/param_tut.php). These are called major stacks parameters and depending on the nature of the experiments such parameters needs to be adjusted wisely.

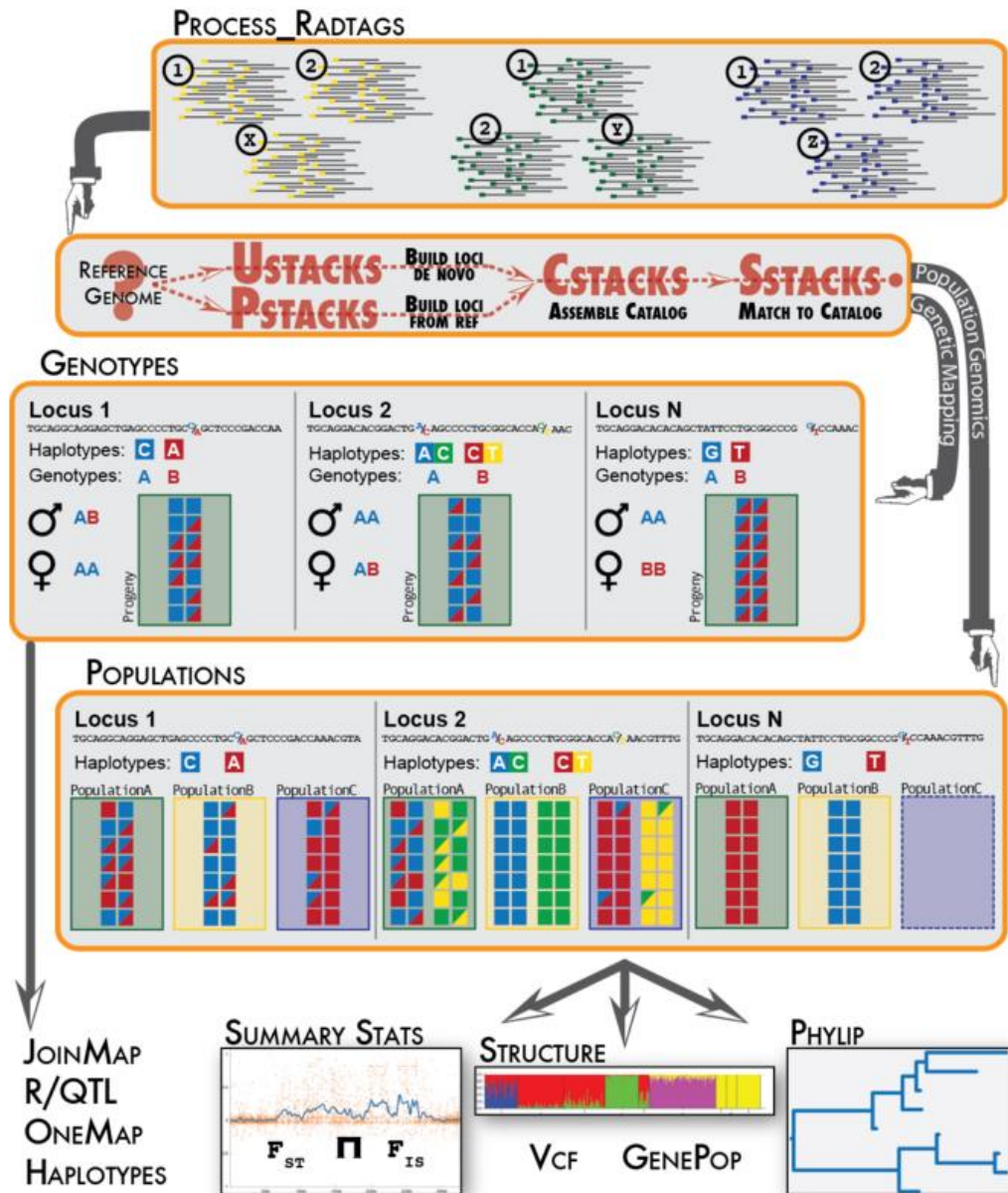


Figure 2.13: A schematic diagram explaining the entire Stacks pipeline in two consecutive stages: *denovo_map.pl* and *ref_map.pl* pipelines for loci building and SNP calling starting from cleaning and demultiplexing through *process_radtags* module. Adopted from Stacks manual (<http://catchenlab.life.illinois.edu/stacks/manual/>).

2.6 Microsatellites

2.6.1 General information

Microsatellite genotyping was carried out with a panel of 11 fluorescently-labelled loci (Ch 2.2, Appendix) in Atlantic salmon (Chapter 4) as a means of double-checking ddRAD sequencing results where varying levels of sire contribution were observed in five of the putative clonal families (G2). The microsatellite loci (Vasemägi et al., 2005) were selected initially for their wide usage and reliability for another project which aims to screen Atlantic salmon broodstock for genetic variation within the same institute.

2.6.2 Fluorescent primer tailing

Polymerase chain reactions were performed using a fluorescent labelled tailed primer method (Boutin-Ganache et al., 2001). The reason behind this was to reduce the cost of purchasing individual fluorescent tagged primers which constitutes the biggest cost for a project involving increasing numbers of markers. The rationale behind the tailed primer method is quite simple: incorporating fluorescent dye into the PCR product to be detected by capillary sequencer. The method employs a two part primer in which a standard primer sequence or *tail* (5'Dye- GGATAACAATTTACACAGG-3') is added to the 5' prime end of the primer sequence. The tail sequence usually corresponds to a readily available standard primer such as an M13 universal primer, meaning the tail on primer should have the same sequence as M13 labelled primer in this case. One of the advantages of this method is that only one primer used in the PCR reaction needs to have the tail, and can be either the forward or the reverse one. Following tail sequence adding, the PCR is performed as standard. The only difference

will be the shift on product range due to the tail sequence compared to the products amplified by untailed primers (in the MA13 example here, product was expected to be 20bp longer due to dye sequence).

In total, three types of dyes (MA13_blue [5'Dye- GGATAACAATTTACACAGG-3'], CAG_green [5'Dye- CAGTCGGGCGTCATCA-3'] and Goddle_black [5'Dye- CATCGCTGATTCGCACAT-3']) were used for 11 microsatellite loci under investigation for the purpose of multiplexing.

2.6.3 Polymerase Chain Reaction (PCR)

Reactions consisted of a total volume of 10 μ L, comprising half volume coming from MyTag™ 2x Mastermix (Bioline, UK) solution ensuring 1X concentration reaction at the end, 3.2 μ L distilled water, 1 μ L DNA (25ng/ μ L) 0.2 μ L of 1uM tailed forward primer, 0.3 μ L of 10uM of non-tailed reverse primer and 0.3 μ L of 10uM fluorescent dye. PCR reactions were conducted on a Biometra TGradient thermal cycler that was programmed with a 1 minute denaturation step at 95°C, 95°C denaturation for 15 seconds; 60°C annealing for 20s and 72°C extension for 30 s for 32 cycles, without requiring a final extension for all PCR multiplexes 1 - 3 (Cp 2.2 Appendix).

2.6.4 Genotyping

Size determination of the fluorescently labelled PCR products was assessed using a Beckman-Coulter CEQ8000 sequencer and associated software. For each capillary run, depending on the fluorescent dye used (0.55 μ L for M13A_blue dye, 0.75 μ L for CAG_green dye and 1.2 μ L for Goddle_black dye), a total volume of 0.55-1.2 μ L of PCR products were added into a 96 well sequence plate with V bottom (Beckman Coulter®, USA) to 30 μ L SLS and 0.4 μ L DNA Size Standard kit-600 (SS600,

Beckman Coulter®, USA) for multiplex_1 and to 30 µL SLS and 0.4µL DNA Size Standard kit-400 (SS400, Beckman Coulter®, USA) for multiplex_2 and multiplex_3. One drop of mineral oil was added at the top of each sample to avoid potential evaporation. An electrophoresis buffer tray, 96 well plate, corresponding the number of sample plate in use, with flat bottom (Beckman Coulter®, USA), was prepared. Each row of 8 samples ran for 45 min using Beckman Frag-3 (size range 60-400bp) genotyping method for multiplex_2 and multiplex_3 and Frag-4 (size range 60-600bp) genotyping method for multiplex_1. Once the run was completed the data was transferred into another computer, licensed software installed, in CEQ format and analyses via Fragment Analysis Module. First, results were viewed in stacked version to identify outlying peaks and confirm reproducibility among samples. Then through single sample view, allele scores were edited manually on annotation editor. Finally, allele scores were extracted and transferred to an Excel spreadsheet as genotypes.

Chapter 3

Multilocus analysis of a meiotic gynogenetic family of European seabass genome using ddRADseq

Münevver Oral^{1†}, Julie Colléter^{2,3†}, Michaël Bekaert^{1†}, John B Taggart¹, Christos Palaiokostas¹, Brendan J McAndrew¹, Marc Vandeputte^{3,4}, Béatrice Chatain³, Heiner Kuhl⁵, Richard Reinhardt⁵, Stefano Peruzzi⁶ and David J Penman^{1*}

¹ Institute of Aquaculture, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK

² Cirad, Persyst, UMR Intrepid, Campus International de Baillarguet, 34398 Montpellier, France

³ Ifremer, Laboratoire de Recherche Piscicole en Méditerranée, Station Expérimentale d'Aquaculture, 34250 Palavas-Les-Flots, France

⁴ Institut National de la Recherche Agronomique, UMR1313 Animal Genetics and Integrative Biology, 78350 Jouy-en-Josas, France

⁵ Max Planck Institute for Molecular Genetics, Ihnestr. 63, 14195 Berlin, Germany

⁶ Department of Arctic and Marine Biology, Faculty of Biosciences, Fisheries and Economics, University of Tromsø, 9037, Breivika, Tromsø, Norway

† Equal contributors

Author contribution: The first draft of the present manuscript was compiled and written in full by the candidate, who was also fully involved in all subsequent revisions. DNA extraction, preparation of RAD libraries (under guidance of JBT), general statistics, data analysis, marker-centromere mapping, construction of the genetic linkage maps and determining cross-over hot spots were carried by the candidate. Bioinformatic part of the study was performed by MB including aligning RAD tags in to genotypes, SNPs calling, final physical location and genetic linkage maps visualising on his program, after marker order provided by MO. Meiotic gynogenetic family for linkage mapping was produced by JC. The other co-authors contributed towards the experimental design, laboratory procedures, in the analysis of the sequenced reads and in the application of the machine-learning algorithm.

Abstract

European Seabass (*Dicentrarchus labrax*) is prime importance specie in Mediterranean countries. Yet, isogenic clonal lines which offer great potential for aquaculture related research have not been established in the species. Production of such lines is trick mainly due to reduced survival and the spontaneous arise of meiotic gynogenetic with varying level of heterozygosity during the production of mitotic gynogenetics. Previous works involved handful of microsatellite and/or morphometric markers to verify the isogeny in the first generation mitotic gynogenetic fish however, increasing number of reliable markers are needed. Double-digest Restriction-site Associated DNA-ddRADseq was used in a single family of 79 offspring from meiotic gynogenetic *D. labrax*, in order to identify SNPs and map female heterogametic markers, particularly those that are at distal end of the chromosome with higher levels of recombination. In total, 54 million raw reads produced with 6,866 unique RAD-tags. A linkage map was constructed based on 764 SNPs that were grouped in 24 linkage groups ($2n = 48$) with a total length of 1,252.02 cM. Physical position of female heterogametic markers as well as microsatellites that are commonly in use were positioned. Recombination frequencies mapped across the genome revealed 0.98 ± 0.12 crossover per chromosome arm, provides evidence to the existing literature in the high levels of interference in fish genomes. Overall the results from this study identifies high number of SNP markers, first time in meiotic gynogenetic family of *D. labrax*, that can be used to overcome one of the bottleneck of producing clonal lines by differentiating meiotic to mitotic gynogenetics, a step in development of clonal lines, needs to be detected and eliminated, to aid the reliable production of isogenic clonal lines. This should also help to speed up the inclusion of isogenic G1 fish derived from many individuals to reveal

genetic variation for many traits. Additionally, this study highlights some parts of genome assembly to be revisited for ordering small contigs.

Keywords: *Dicentrarchus labrax*, Meiotic gynogenesis, Clonal lines, ddRAD seq, Genetic map, Aquaculture.

3.1 Introduction

Chromosome set manipulation is a methodology that has been exploited over a long period in fish research (Purdom, 1983; Thorgaard 1983; Ihssen et al., 1990; Hulata, 2001; Gomelsky, 2003). The ability to retain the second polar body post-fertilisation and / or suppress other early cell divisions by temperature, chemical or pressure shocks, coupled with the relative ease of gamete inactivation by irradiation has led to its widespread use. The various chromosome sets that can be generated (e.g. haploids, triploids, tetraploids, androgenetics, meiotic or mitotic gynogenetics) have been exploited in a wide range of studies including gene mapping (Danzmann & Gharbi, 2001; Nichols et al., 2003), genome assembly (Brawand et al., 2014; Lien et al., 2016), construction of isogenic clonal lines (Bongers et al., 1998; Muller-Belecke & Horstgen-Schwark, 2000) and production of sterile farm fish (Chourrout & Quillet 1982; Preston et al., 2013).

Though widely practised, there are a number of technical pitfalls that can impact the effectiveness of chromosome set manipulation procedures. For example, there can be a potential genetic contribution from the irradiated gamete source, this being associated with poorly optimised protocols leading to incomplete ablation (Komen & Thorgaard, 2007). Furthermore, the efficacy of protocols designed to retain chromosome sets post fertilisation / activation can also be severely affected by gamete quality and slight

alterations in the timing and intensity of the applied shock (Yamamoto, 1999; Kato et al., 2002; Bertotto et al., 2005). Spontaneous retention of the second polar body (Braasch & Postlethwait, 2012; Havelka et al., 2016) may also generate additional unexpected (and unwanted) ploidy states.

Throughout the development of the technology, genetic markers have been used to monitor the effectiveness of the procedure. To date this has generally involved screening with a small panel of available markers, to confirm the presence / absence of particular parental chromosomal sets. These markers include pigmentation genes, allozymes, multilocus minisatellites and microsatellites (Komen & Thorgaard, 2007). While this approach can give a broad indication as to the effectiveness of the treatment, it is relatively insensitive for detection and quantification of potential instances of aneuploidy. Another limitation to using a small number of markers is that those that happen to be located close to centromeric regions will be compromised with respect to their ability to detect crossover events. This is a key requirement, for example, for differentiating between mitotic and meiotic gynogenetics; i.e. informative telomeric markers will be heterozygous in meiotic gynogenetics and homozygous in mitotic gynogenetics, while centromeric markers will largely be homozygous in both types. For most studies to date marker-centromere distances have been unknown.

The advent of genotyping by sequencing approaches that exploit next generation sequencing technologies (Davey et al., 2011) permits the simultaneous discovery and screening of large numbers of single nucleotide polymorphisms (SNPs) per individual at a realistic cost. This provides an opportunity to more accurately assess the effectiveness of various elements of chromosomal set manipulation procedures. In this study SNPs generated by double digest restriction-site associated DNA (ddRAD) sequencing (ddRAD seq; Peterson et al., 2012) were employed to comprehensively

examine parental genetic contributions in an experimentally generated meiotic gynogenetic family of European seabass *Dicentrarchus labrax*. The main objectives of the study were to (i) look for potential paternal contribution from UV-irradiated sperm; (ii) generate a SNP locus - centromere map alongside with (iii) a genetic linkage map based on meiotic gynogenetic family; and (iv) screen informative (female heterozygous) markers for their potential to distinguish between mitotic and meiotic gynogenetics.

3.2 Materials and Methods

3.2.1 Production of mapping family - Meiogynogenetics

The meiotic gynogenetic seabass family was produced at the Ifremer Experimental Aquaculture Station (Palavas-les-Flots, France), using parent fish from a West Mediterranean broodstock population. Broodstock were aged 4 to 6 years and weighed 1 to 5 kg, and were kept in recirculating systems (8 m³ tanks, rate of O₂ enriched water renewal 250 Lh⁻¹, constant low aeration) maintained under natural conditions of temperature and photoperiod (43° 31' 40 N, 3° 55' 37 E) and fed commercial diets (NeoRepro, Le Gouessant, France). Spermiating males were identified by gentle abdominal pressure and held in a handling tank. Female maturation stage was assessed from ovarian biopsies obtained by introducing a thin catheter (Pipelle de Cornier, Laboratoire CCD, Paris, France) into the genital orifice. Females at the correct stage of development received a single dose (10 µg.kg⁻¹) of Luteinizing Hormone Releasing Hormone analogue (LHRHa, Sigma, France) in order to induce final maturation and ovulation. The UV irradiation device, used to inactivate the paternal genome, comprised of eight UV lamps (12 W, 254 nm, Vilber-Lourmat, Marne-la-Vallée, France) fixed above and below (four lamps each) a quartz plate which was mechanically agitated to

stir sperm samples throughout irradiation. Diluted sperm (5 mL) from a single male (diluted 1:20, v/v in artificial extender Seabass Gamete Short term Storage made of Storefish (IMV Technologies, France)) was irradiated in an 8.5 cm diameter quartz petri dish for 8 minutes to apply a total dose of 326 mJ/cm² (Peruzzi & Chatain, 2000). The irradiated sperm were added to 125 mL of eggs (untreated, good quality) and then 125 mL of seawater was added to initiate fertilisation. A pressure shock of 8500 psi and 4 min duration was applied, starting at 6 min after fertilisation, to restore diploidy via retention of the second polar body. All procedures were performed under total darkness in a temperature-controlled room maintained at 14°C. Eggs were incubated in 40 L tanks in a dedicated recirculating water system (temperature 14-14.5°C; salinity 35-36‰) until hatching. All tanks were maintained in darkness until sampling. Ten days after hatching, a subset of 80 larvae were fixed in 99% ethanol; fin tissue from parents was also fixed in ethanol.

3.2.2 DNA preparation

DNA was extracted from all 80 offspring (entire larva) and both parents (fin tissue) using a commercial salting out kit (REALpure DNA extraction kit; REAL Laboratories, Durviz, Spain) according to the manufacturer's protocol. This included the recommended RNase incubation step to reduce RNA contamination in the final product. The DNA concentration and purity of each sample was assessed by spectrophotometry (Nanodrop, Thermo Scientific, UK), while its integrity was assessed by 0.7% agarose gel electrophoresis. Each sample was then preliminarily diluted to c. 50 ng/μL in 5 mM Tris, pH 8.5. A final, more accurate, fluorometric-based assessment of DNA concentration was then performed on all samples using the Qubit® dsDNA HS Assay Kit (Invitrogen, UK). Fluorescence measurements (20 uL volumes) were performed on

a 96 well qPCR thermal cycler (Quanta, Techne, UK), with seabass DNA concentrations being derived from a calibration curve generated from a set of standard dsDNAs. Based on these readings the seabass samples were diluted to c. 10 ng/ μ L in 5 mM Tris, pH 8.5 for use in ddRAD library construction protocol.

3.2.3 ddRAD library preparation and sequencing

The ddRAD library preparation protocol used here is described in detail elsewhere (Manousaki et al., 2015; Brown et al., 2016). Briefly, a single restriction enzyme digestion / adapter ligation reaction was performed for each progeny sample, while triplicate reactions were made for both dam and sire DNA samples. The latter ensured high coverage in parental samples in order to more confidently assign true SNPs in the pedigree. Each sample (40 ng DNA) was digested at 37°C for 30 minutes with 0.8 U *Sbf*I ('rare' cutter, CCTGCA|GG motif) and 0.8 U *Sph*I ('common' cutter, GCATG|C motif) high fidelity restriction enzymes (New England Biolabs; NEB) in a 6 μ L reaction volume that included 1 \times CutSmartTM buffer (NEB). After cooling the reactions to room temperature, 3 μ L of a premade barcode-adapter mix was added to the digested DNA, and incubated at room temperature for 10 min. This adapter mix comprised individual-specific barcoded combinations of P1 (*Sbf*I-compatible) and P2 (*Sph*I-compatible) adapters at 6 nM and 72 nM concentrations respectively, in 1 \times reaction buffer 2 (NEB). Adapters were compatible with Illumina sequencing chemistry (see Peterson et al. 2012 for details). The barcoded adapters were designed such that adapter-genomic DNA ligations did not reconstitute RE sites, while residual RE activity limited concatemerization of genomic fragments during ligation. The adapters included an inline five- or seven-base barcode for sample identification (Table S1, Appendix). Ligation was performed over 40 min at 22°C by addition of a further 3 μ L of a ligation

mix comprising 4 mM rATP (Promega, UK), and 2000 cohesive-end units of T4 ligase (NEB) in 1× CutSmart buffer.

The ligated samples were then heat denatured at 65°C for 20 min, cooled, and combined into a single pool. The pooled sample was column-purified (MinElute PCR Purification Kit, Qiagen, UK) and size selection of fragments, c. 320 bp to 590 bp, was performed by agarose gel electrophoresis. Following gel purification (MinElute Gel Extraction Kit, Qiagen, UK) the eluted size-selected template DNA (60 µL in EB buffer) was PCR amplified (11 cycles PCR; 28 separate 12.5 µL reactions, each with 1 µL template DNA) using a high fidelity Taq polymerase (Q5 Hot Start High-Fidelity DNA Polymerase, NEB). The PCR reactions were combined (350 µL total), and column-purified (MinElute PCR Purification Kit). The 55 µL eluate, in EB buffer, was then subjected to a further size-selection clean-up using an equal volume of the AMPure magnetic beads (Perkin-Elmer, UK), to maximize removal of small fragments (less than ca. 200 bp).

The final library was eluted in 20 µL EB buffer and sequenced over two full Illumina MiSeq runs (v2 chemistry, 300 cycle kit, 162 bp paired end reads; Illumina, Cambridge, UK; 10.5 pM library applied and both runs spiked with 3% Illumina phiX control DNA). The raw sequence data from this study were deposited at the EBI Sequence Read Archive (SRA) with the accession number ERP006697.

3.3 Data Analysis

3.3.1 Genotyping ddRAD alleles

Following initial analysis (FastQC: Andrews, 2010) to confirm that high-quality sequence data had been generated, the MiSeq reads were processed using Stacks (v.1.

17; Catchen et al. 2013), a package designed specifically to identify and score SNPs from restriction-enzyme based sequence data. First, the ‘process_radtags’ function was used to demultiplex the individual samples. During this process sequence reads with quality scores below 20 (-s set to 20), missing either restriction site or with ambiguous barcodes were discarded. Barcodes were removed and all sequences were 3’ end trimmed to be 148 bases long. Then reference based Stacks analysis was performed, using ‘ref_map.pl’ perl script. Sequence alignment/ map (SAM) files were created using Bowtie aligner (Langmead & Salzberg, 2013) and a pre-release version of the seabass genome (since published; Tine et al. 2014). The main Stacks parameter values used in this analysis were $m = 10$ and $n = 1$. In order to maximise the number of informative markers investigated while minimising missing or erroneous data, only polymorphic ddRAD-tags that containing 3 or less SNPs (maximum of 4 alleles) and which were detected in both parents and present in at least 75% of the offspring were scored.

3.3.2 Genetic linkage map construction

It was not feasible to construct a genetic linkage map *de novo* from unordered meiotic gynogenetic family data. Both R/OneMap (Margarido et al., 2007) and TMAP (Cartwright et al., 2007) were explored for genetic linkage map construction without success. The final map was constructed using R/OneMap after assigning markers to linkage groups based on the seabass genome assembly (see section 3.4.3). Genotypes were imported in *outcross* format into R/OneMap in a modified way such that all genotypes shared the same segregation pattern (“ab x ab cross”). This package uses Hidden Markov Models (HMM) algorithms for outbred species while in parallel implements the methodology described in Wu et al. (2002), for calculating the most probable linkage phase. Recombination fraction between all pairs of markers was

calculated using *rf.2pts* function. These groups were ordered using the *order.seq* function in four available two-point based algorithms including *ser*, *rcd*, *rec* and *ug* and the one which gave the smallest distance was selected for each LG. Following ordering, markers in the same LG were forced to the final map by using *force* function after inspection of *safe* order. The order of markers was also inspected visually using *rf.graph.table* which plots a heat map of LOD score and recombination frequency. Map distances were calculated in centiMorgans (cM). Genetic Mapperv0.3 (Bekaert, 2012) was used for the final visualisation of genetic linkage map of meiotic gynogenetic *D.labrax*.

3.3.3 Visualising physical position of markers and microsats from previous studies

The output of genome aligner (SAM files) were used for the positioning each ddRADseq locus and visualised using Genetic-Mapper v0.5 (Bekaert, 2015). Eleven microsatellite markers (García De León et al., 1995; Chistiakov et al., 2005) that have been used to differentiate between meiotic and mitotic gynogenetic sea bass (Colléter, 2015) were also assigned to the physical map once the genomic positions in basepairs were identified using Blastn ($1E^{-20}$ and lower).

3.3.4 Marker-centromere mapping

Centromeres are expected to be in regions with zero or low heterozygote frequency, with an increase in heterozygote frequency towards the telomeres. For each maternally informative ddRADseq locus, heterozygosity (y) was computed across all progeny. Marker-centromere map distances (in cM) were calculated using the formula $100*(y/2)$,

under the assumption of complete interference, believed to be characteristic of fish species (Thorgaard 1983; Sakamoto et al., 2000; Nomura et al., 2006).

3.3.5 Comparison of genomic assembly with linkage maps

The genome assembly and the linkage map generated in the present study were compared to the recently published high-density SNP-based genetic linkage map of Palaiokostas et al. (2015b), as an independent source for comparing marker order. For comparison with the genome assembly, the loci from Palaiokostas et al. (2015b) were physically located as described above (section 3.3.1). Common polymorphic loci between the two linkage maps were identified by BLASTn. First, the loci including the common enzyme recognition site (“TGCA”; *Sbf*I) from the present study (in total: 395 markers out of 764 female heterogametic assigned markers) were trimmed down to 95bp, compatible with the RADseq P1 read length of Palaiokostas et al. (2015b). Then a local nucleotide database was generated on Bioedit (version 7.2.5) (Hall, 1999) from all assigned markers of Palaiokostas et al. (2015b) and all polymorphic markers of the present study were blasted against them. Stringent filtering options were applied to tabular output based on: i) e^{-20} and smaller; ii) alignment length 90bp and higher (which ensures 94.7% similarity rate); and iii) sequences with more than 10 mismatches were removed from the dataset.

3.4 Results

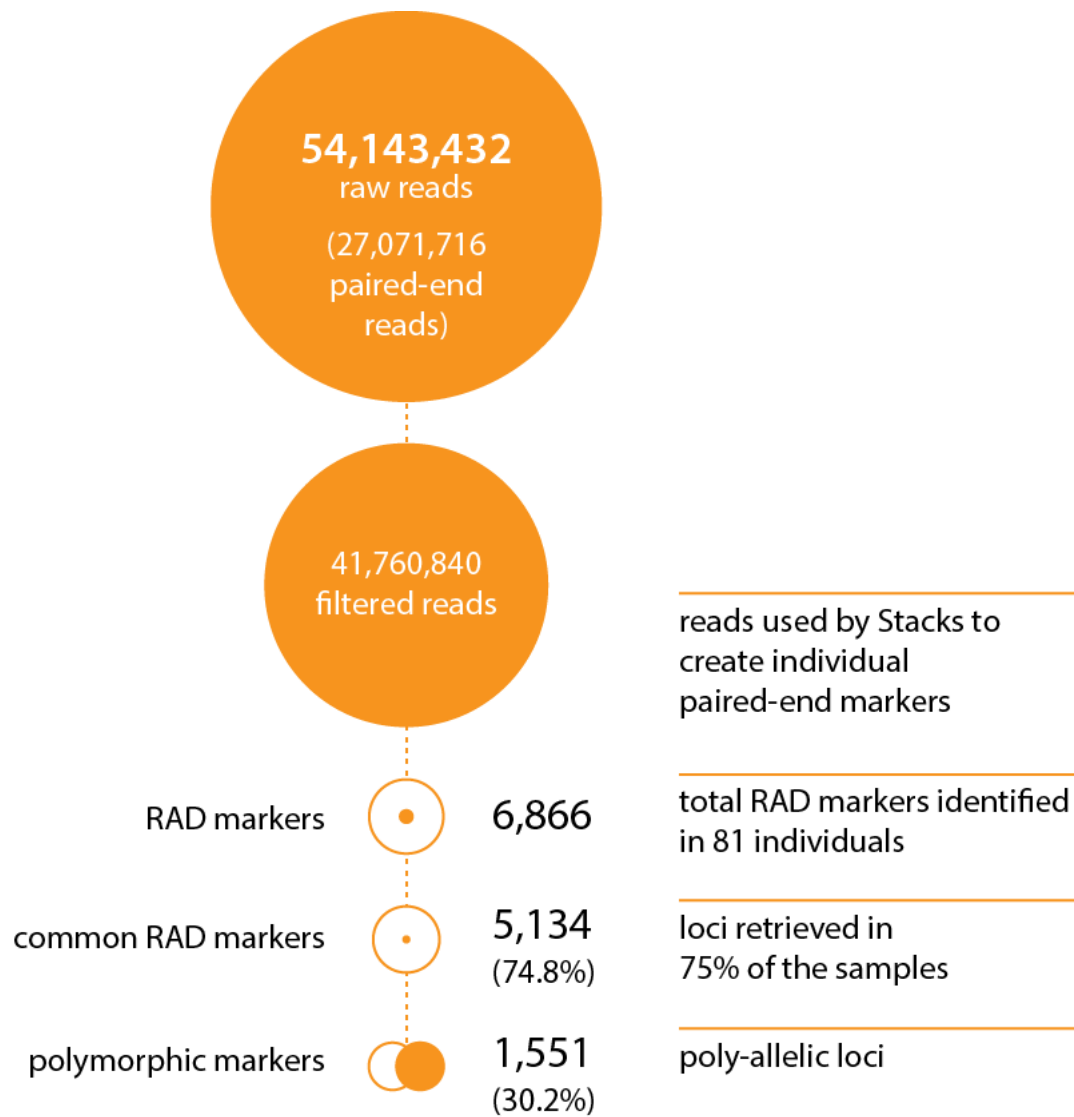


Figure 3.1: Sequencing and ddRAD-tag summary. Detailed number of reads before and the after filters (orange disk) followed by the reconstructed numbers of ddRAD markers and polymorphic ddRAD markers (orange circles).

3.4.1 ddRAD sequencing

A total of 27,071,716 paired-end raw reads were produced from the combined two sequencing runs for the meiotic gynogenetic *D. labrax* family (see Fig 3.1; detailed information for each sample used is provided in S1 Table, available in electronic version). Following demultiplexing using *process_radtags*, 77.1% of the raw paired-end reads were retained (20,880,420). Only one sample offspring (MO241) failed to produce sufficient reads (c.1542 reads < 150 K) and was dropped from subsequent STACKS analyses. As planned, the read numbers for both parents (785 K, sire & 1127 K, dam) exceeded those of offspring by a factor of c. 2 (average no. per reads per offspring, 504 K). Read numbers for each sample are detailed in Table S2 (available in electronic version). The reference-based Stacks analysis identified 6,886 unique ddRAD loci and 1,551 potential SNP loci.

3.4.2 Investigation of potential sire contribution

Within the polymorphic marker dataset 340 SNPs were identified with male informative alleles, i.e. one (214 loci) or both (126 loci) alleles at a locus detected in the male parent alone. No male-specific alleles were detected in any of the offspring. Later mapping of these loci to the seabass reference genome confirmed that these markers were located across all seabass chromosomes. Thus no sire contribution was detected within the ddRAD dataset for this gynogenetic family.

3.4.3 Construction of female genetic linkage map

With the absence of paternal alleles confirmed, the marker dataset was refined to produce a robust set of informative SNPs for female map construction. Dam homozygous markers were removed (non-informative: 687 loci) as were loci where the minor allele frequency was <0.4 among the progeny samples (8 loci). Additionally 52

loci were removed since both parental genotypes were missing. This left data from 804 female-informative SNPs to be used in linkage map construction.

The linkage map (constructed using a LOD score of 4-5) comprised 764 ordered SNPs and was 1,252 cM in length (Figure 3.2; Table 3.1). Average marker distance was 1.63 cM. Linkage groups were between 23 cM (LG 3) and 78 cM (LG 1A) in length (mean 52 cM) and comprised between 15 (LG 18-21) and 46 markers (LG20; mean 32; see Dataset S1 for the sequence of all markers assigned, available in electronic version). As the initial grouping of SNPs within the linkage map was based on the genome assembly, the distribution of markers was in accordance with 24 chromosomes in *D. labrax*.

Table 3.1. Meio gynogenetic *D. labrax* genetic linkage map

LGs	No. of Markers	Size (cM)
LG 1A	45	78.04
LG 1B	29	51.30
LG 2	30	61.83
LG 3	24	22.79
LG 4	34	44.10
LG 5	38	68.19
LG 6	26	55.59
LG 7	26	72.31
LG 8	31	47.92
LG 9	23	46.00
LG 10	30	34.68
LG 11	42	54.26
LG 12	29	47.25
LG 13	27	54.03
LG 14	31	66.67
LG 15	37	61.31
LG 16	37	49.89
LG 17	45	55.03
LG 18-21	15	30.23
LG 19	30	56.96
LG 20	46	45.82
LG 22-25	39	62.70
LG 24	19	39.07
(LG X)	31	45.05
Total	764	1252.02

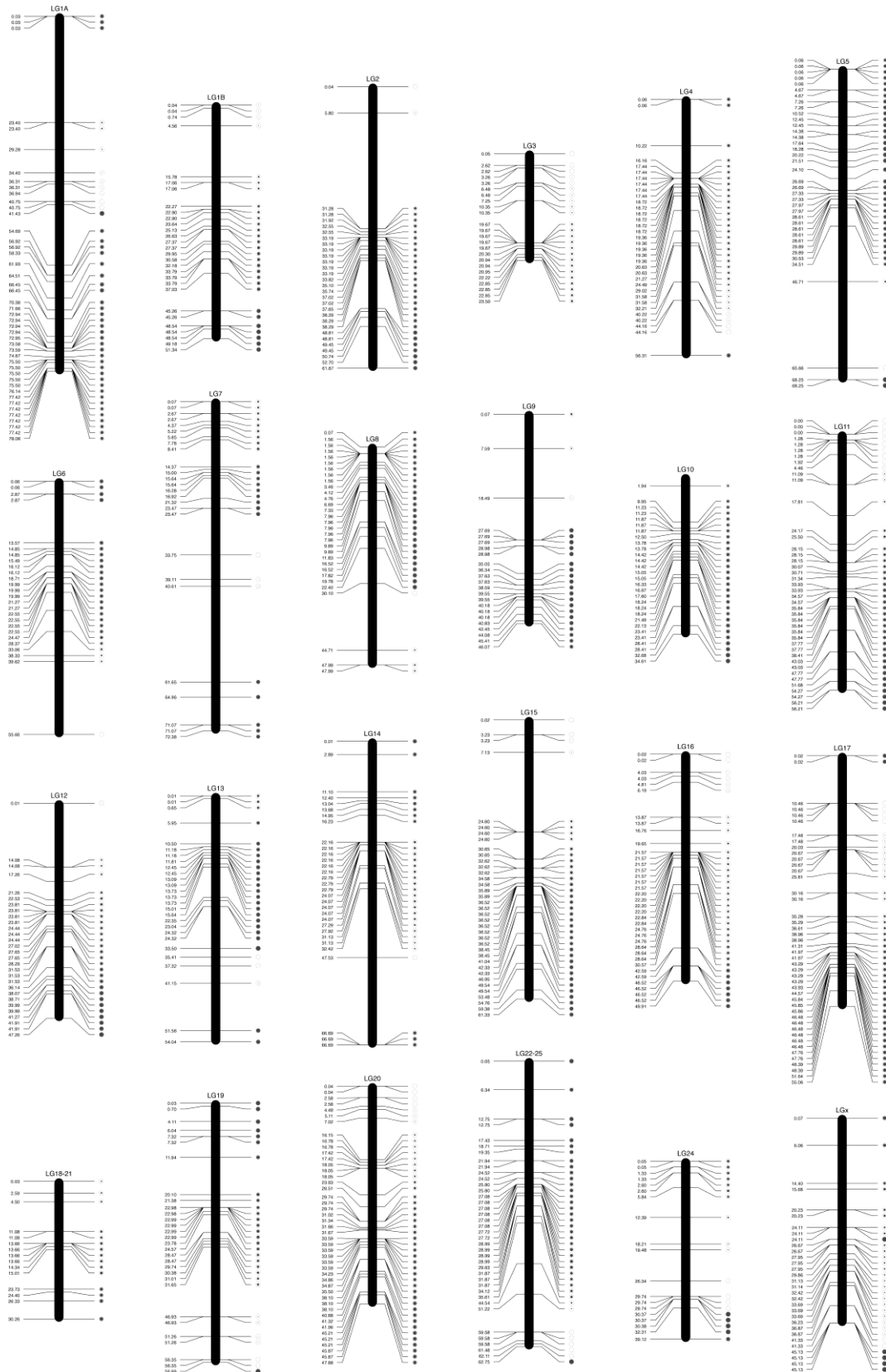


Figure 3.2: Genetic linkage map of meiotic gynogenetic *D. labrax*. The positions on the left side of chromosomes are the distance in centiMorgans (cM), the circles on the right hand side represent observed heterozygosity levels at each map position (empty circles represent homozygotes whereas increasingly filled black dots represents the higher levels of heterozygosity). Detailed data are provided in Table S3.

3.4.4 Physical position of markers in seabass genome

Figure 3.3 demonstrates the visual representation of the markers on the genome assembly, covering all 24 chromosomes in the genome of *D.labrax*. The position of each marker in the genome assembly is shown in Table S3 (marker ID and physical position is represented per chromosome, available in electronic version).

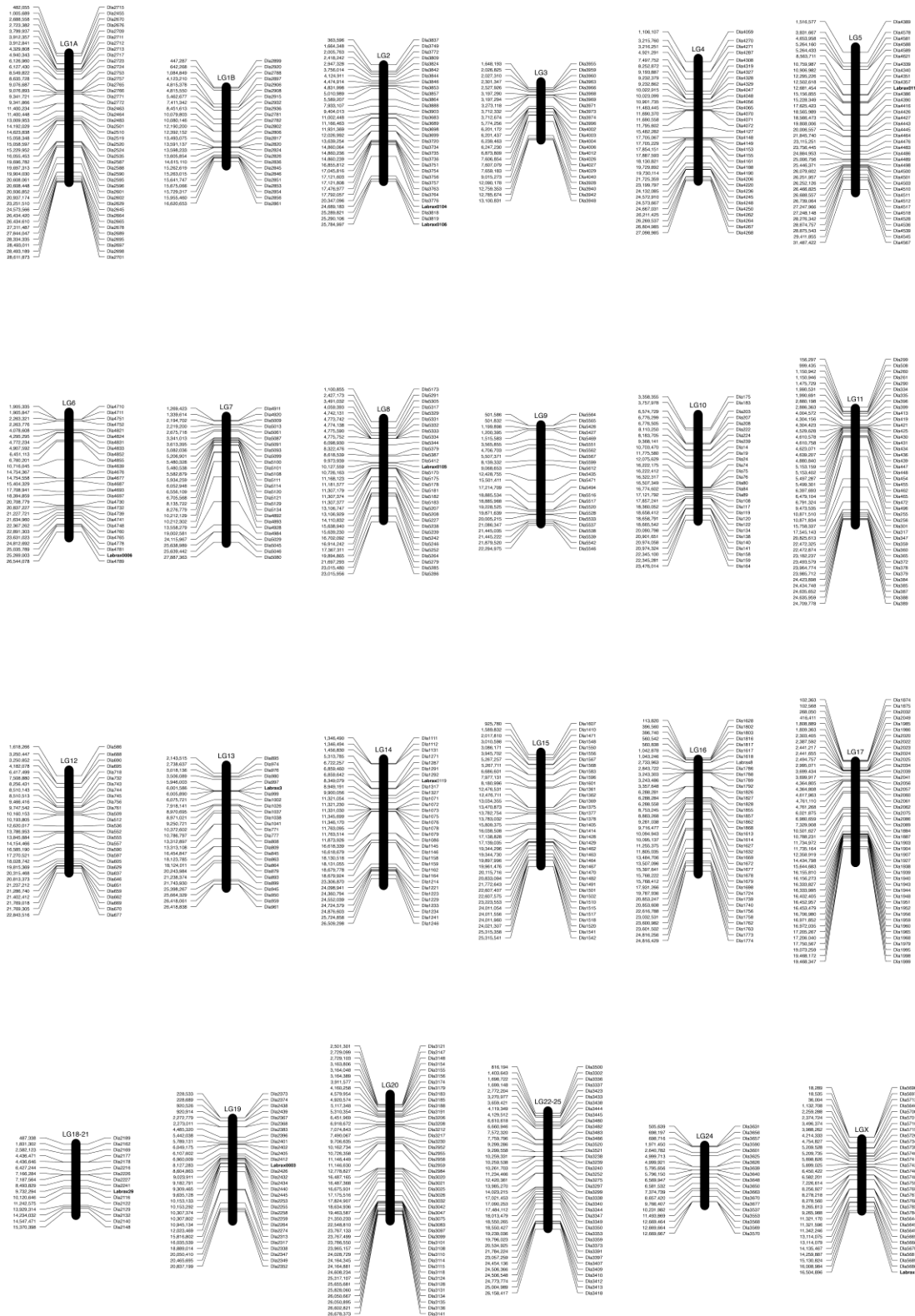


Figure 3.3: Physical map position of SNP markers that have been identified in the present study from meiotic gynogenetic *D. labrax*. The positions on the left side of chromosomes represent physical map positions in basepairs while marker IDs are given on the right side of the chromosomes. Detailed data are provided in Table S3.

3.4.5 Marker-centromere mapping

Estimated M-C recombination rates ranged between zero and one (i.e. 0 to 50 cM map distances under the assumption of complete interference, existence of one crossover decreases the possibility of having a second crossover nearby). Fig 3.8 shows a histogram of recombination frequencies and Table S6 shows marker-centromere map distances, available in electronic version. Crossover frequencies were detected as shown in Fig 3.4. Seven loci (0.87% of total loci) showed 100% recombination (i.e. telomeric), while 16 loci (1.99%) showed zero recombination (i.e. centromeric). Almost half of the markers had heterozygote frequencies above 0.667 (49.12%), the expected maximum theoretical value for independent segregation between a marker and the centromere due to multiple crossovers, indicating high interference.

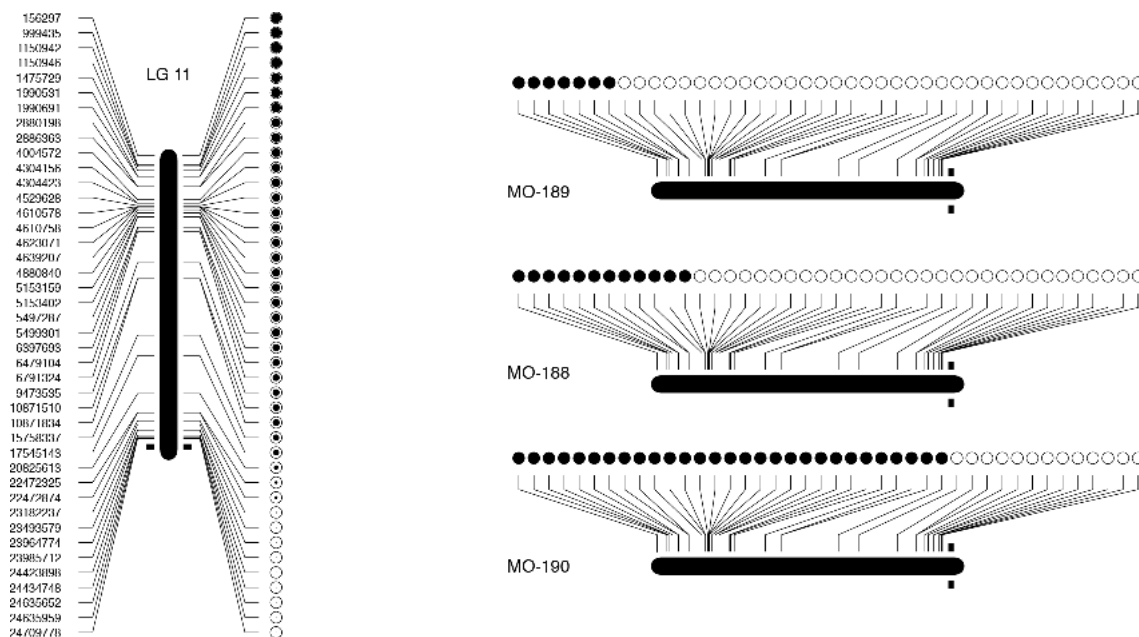


Figure 3.4: Detailed example of mapping in a single seabass linkage group (LG 11), illustrating the computed recombination fraction for 79 progeny. Empty circles represent homozygotes close to the centromere (represented by black boxes either side of the linkage group), and increasing black dots represent higher levels of scored heterozygotes towards the telomeric region. The panel to the right represent randomly chosen individuals from the meiotic gynogenetics family, showing the recombination points in LG 11.

Eleven chromosomes (LG 1B, 2, 3, 6, 10, 11, 12, 15, 16, 18-21 and 20) showed single armed (mono-armed) behaviour, with heterozygosity rising from one end of the

chromosome to the other reaching up to almost 100% (see Fig 3.5 as an example of monoarm chromosome correlation graph). Three chromosomes (LG 4, 19 and 22-25) fitted the mono-armed pattern with the exception of a single outlying marker (i.e. heterozygosity for one marker did not fit the overall pattern). Three chromosomes (LG 14, 17 and 24) represented a clear bi-armed pattern (intermediate region with very low heterozygote frequency, rising towards a high frequency at either end) (Fig 3.6 represents example of biarmed chromosome graphs). One chromosome (LGX) fitted the bi-armed pattern with the exception of a single outlying marker (i.e. heterozygosity for one marker did not fit the overall pattern). Six chromosomes (LG 1A, 5, 7, 8, 9 and 13) did not show a clear pattern of heterozygosity along the chromosome that could enable us to assign an arm structure (mono-armed or bi-armed). Figure 3.7 show the graphs supporting this in LG7 as an example.

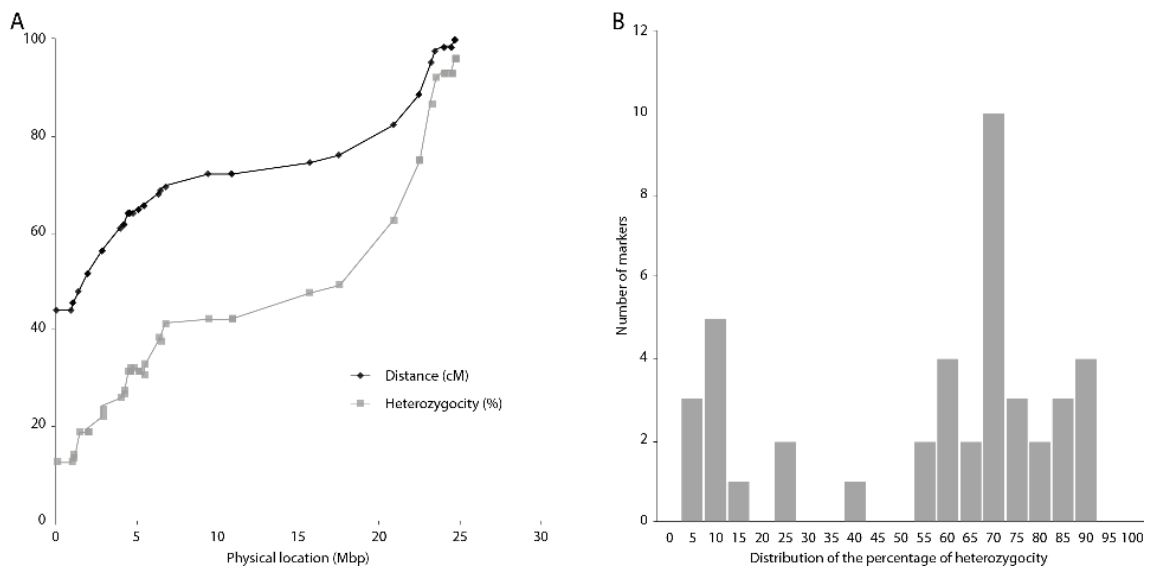


Figure 3.5: Example of mono-arm chromosome, LG 11. A) Correlation of physical location (Mbp) with the percentage heterozygosity and genetic linkage map distance (cM) with genome assembly (Mbp). B) Frequency distribution of markers based on percentage of heterozygosity.

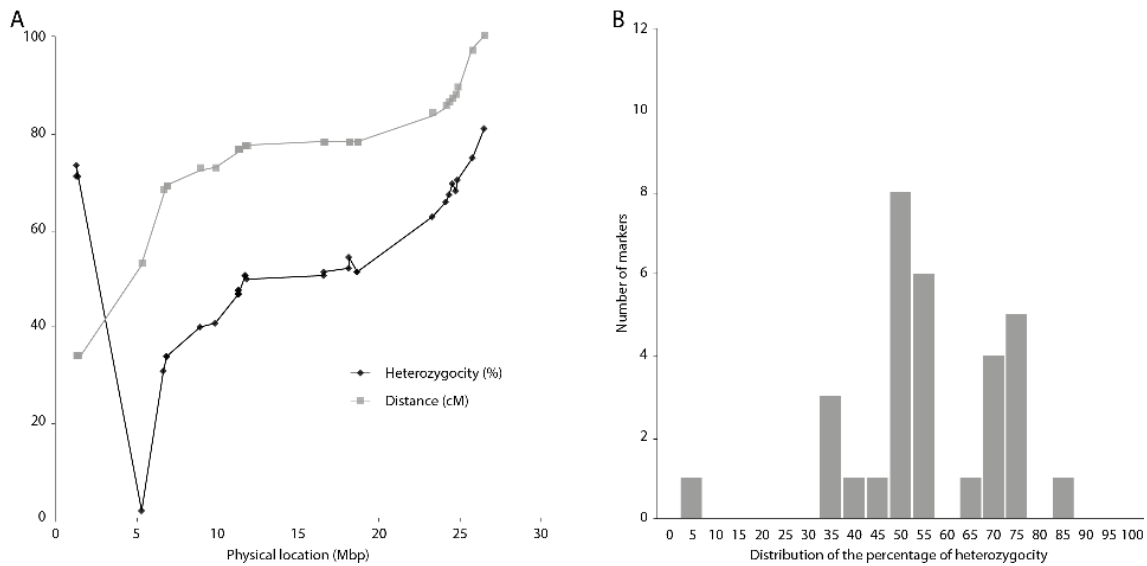


Figure 3.6: Example of bi-arm chromosome, LG 14. A) Correlation of physical location (Mbp) with the percentage heterozygosity and genetic linkage map distance (cM) with genome assembly (Mbp). B) Frequency distribution of markers based on percentage of heterozygosity.

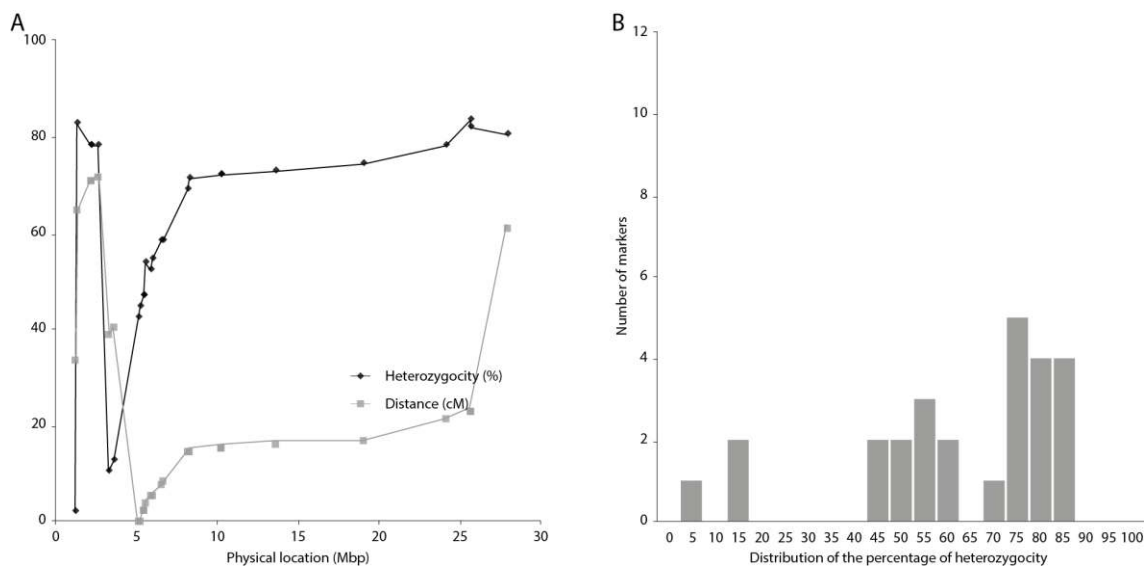


Figure 3.7: Example of ambiguous chromosome, LG 7. A) Correlation of physical location (Mbp) with the percentage heterozygosity and genetic linkage map distance (cM) with genome assembly (Mbp). B) Frequency distribution of markers based on percentage of heterozygosity.

To explore this further, we compared the RAD locus positions from the dense linkage map of Palaiokostas et al. (2015b) with those in the genome assembly. All of the linkage groups of Palaiokostas et al. (2015b) contained markers from the corresponding chromosome in the genome assembly, plus additional markers from unassigned (UNK) scaffolds. There were no cases where markers were assigned to different chromosomes in the assembly. The correlations for each linkage group are shown in Table S7,

available in electronic version. The six LGs which did not show a clear pattern of heterozygosity in the current study were all among the ten LGs showing the lowest correlation in marker order between the dense linkage map and the physical assembly, suggesting an association between the accuracy of the genome assembly and the clarity of arm structure derived from the present data. Of the 764 ddRADseq markers in the linkage map based on the meiotic gynogenetic family, 63 (8.2%) were also found in the RADseq linkage map of Palaiokostas et al. (2015b). All of these were found in the same linkage groups in both maps.

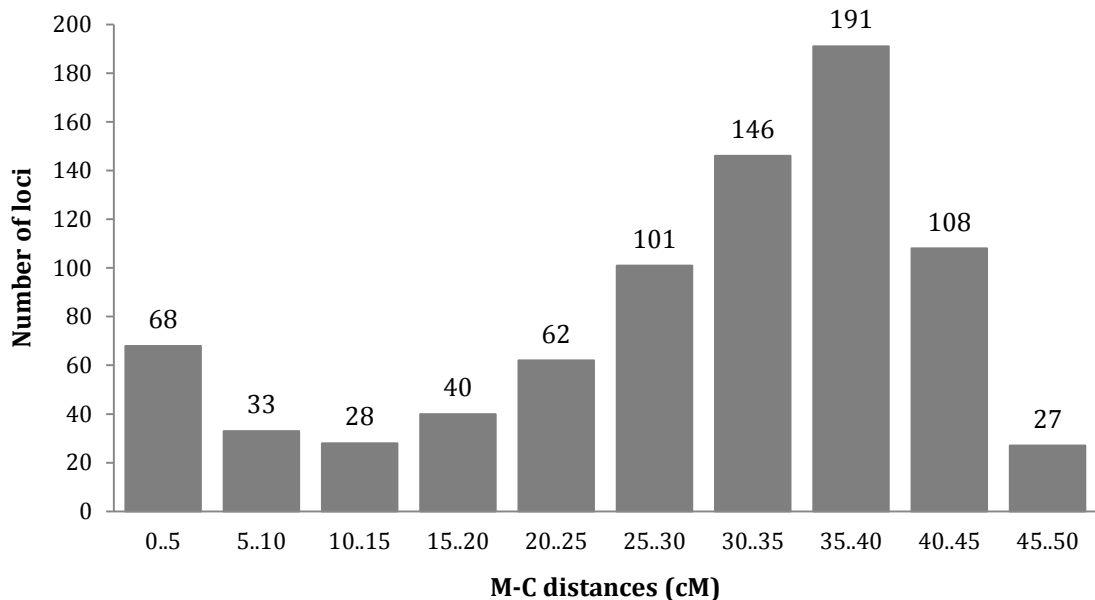


Figure 3.8: Frequency distribution of marker-centromere distances under the assumption of complete interference at 804 female heterogametic loci in meiotic gynogenetic European seabass.

After removing the six chromosomes that did not show clear heterozygosity patterns (LG 1A, 5, 7, 8, 9 and 13) and the single anomalous markers in three chromosomes (LG 4, 19 and 22-25), the mean recombination frequency per chromosome arm was 0.989 (S.E. 0.123). However, there were instances of multiple crossovers in some chromosome arms (Table S5, available in electronic version).

3.5 Discussion

European seabass is an important mariculture species, extensively farmed in the Mediterranean basin. The need to develop genetic and genomic resources to underpin future development of this species is clearly recognised, and has resulted in the production of a first draft genome assembly (Tine et al., 2014), a number of linkage maps (Chistiakov et al 2005, 2008; Palaiokostas et al., 2015b) and a radiation hybrid panel (Guyon et al., 2010). A further key resource would be the development of isogenic clonal lines through androgenesis (Colléter et al., 2014) or mitotic gynogenesis (Bertotto et al., 2005). These have not been successfully established yet despite some efforts (Peruzzi and Chatain, 2000; Francescon et al., 2004, Bertotto et al., 2005; Colléter et al., 2014). One of the bottlenecks in the production process is spontaneous meiotic gynogenetics, which are thought to arise through retention of the second polar body therefore having varying levels of heterozygosity. These need to be detected and eliminated from putative mitotic gynogenetic fish for the reliable production of isogenic clonal lines in the subsequent generation. To address this bottleneck, the present study constructed the first gene-centromere linkage map (of moderate marker density) for this species, in order to identify markers at the distal end of the chromosomes. Such markers are more informative in discriminating between mitotic and meiotic gynogenetics, due to their higher recombination frequencies. This study also explored a second technical issue in the production of gynogenetic fish, that of potential paternal contribution following UV irradiation of sperm, by analysing large numbers of informative SNP markers (compared to smaller numbers of markers in previous studies on fish species). The genotyping-by-sequencing approach used in this study (ddRADseq) proved to be very successful for both objectives, and also to be cost-effective for this purpose, generating 804 female informative markers for the gene-centromere map and 340

informative markers for assessing potential paternal contribution, from the analysis of two subsequent sequencing run of the same ddRADseq library. It is feasible to prepare and sequence such a library in one to two weeks for relatively modest cost, and this technique could thus be used routinely in verifying the development of isogenic clonal lines in this and other fish species. RADseq (Baird et al., 2008) and its derivative ddRADseq (Peterson et al., 2012) have already been used for genetic linkage mapping in model and non-model organisms (Anderson et al., 2012; Recknagel et al., 2013; Gonen et al., 2014; Kai et al., 2014; Palaiokostas et al., 2015a), studies on sex determination systems (Palaiokostas et al., 2013a) and QTL analysis (Houston et al., 2012).

Although 11 microsatellites (Chistiakov et al., 2005; García De León et al., 1995; Colléter, 2015) that are currently in use (positioned on physical map, Fig 3.3) for initial screening were validated in three meiotic gynogenetic family in seabass previously, limited numbers of loci are in use might give rise to false positive identifications (e.g: meiotic gynogenetic fish might be identified as mitotic gynogenetic).

A requisite for successful production of uniparental fish is the ability to completely ablate the genetic material in the irradiated gametes. In this study, 340 male informative SNP markers were identified, none of which were detected in any of the 79 progeny. These markers were located across all 24 linkage groups, confirming a lack of paternal contribution at this level of resolution. It is clear that using this protocol (developed by Peruzzi and Chatain, 2000) we were able to produce a robust gynogenetic family, suitable for gene-centromere mapping.

A gene-centromere map, comprising 764 SNPs spanning 1,252.02 cM with an average marker distance of 1.63 cM, was constructed. Approximately 95% of the female-informative SNPs (764 out of 804) were successfully placed on the linkage map. The

genetic linkage map constructed in the present study was shorter than that produced by Palaiokostas et al., (2015b), which had a total length of 4,816 cM. The length of *D. labrax* linkage groups in the present study varied from 22.79 cM to 78.04 cM and exhibited a positive correlation, in most cases, with the number of markers mapped per linkage group. Marker-centromere frequencies were ranged between 0 and 1 (0 and 50 cM). These results clearly demonstrated that SNP loci produced by ddRAD sequencing were widely distributed in the seabass chromosomes, covering the entire chromosomal regions from centromeric to telomeric locations. Theoretically under the assumption of no interference (with only a single crossover event taking place between non-sister chromatids), the maximum frequency of heterozygotes should be 67% at the telomeres. However out of 804 female heterogametic SNP loci, 395 loci (49.12%) showed heterozygote frequencies above 0.67, indicative of crossover interference in seabass chromosomes. This phenomenon is well documented in the literature for other fish species (Thorgaard, 1983; Danzmann & Gharbi, 2001; Morishima et al., 2001; Nomura et al., 2006). Martínez et al. (2008) observed similar proportion of markers (48.1%) with heterozygosity exceeding 0.67 in turbot (*Scophthalmus maximus*). Twenty-seven of the seabass SNPs showed over 90% heterozygotes in the meiotic gynogenetic family (of which seven showed 100% heterozygotes), suggesting that these could be used in individual SNP assays as a smaller scale assay for discriminating between meiotic and mitotic gynogenetics. At the centromeres of the chromosomes, 68 loci showed less than 10% heterozygotes (of which 16 showed no heterozygotes).

Thorgaard (1983) reported high levels of interference in rainbow trout. Subsequent literature suggests that high crossover interference is a wide-spread phenomenon in fish and shellfish species (Martínez et al., 2008; Morishima et al., 2001; Nie et al., 2012; Reid et al., 2007; Thorgaard, 1983). The results from the present study in general

support this, with an average recombination frequency of around one. After removing the six chromosomes that did not show the clear heterozygosity patterns (LG 1A, 5, 7, 8, 9 and 13) and the single anomalous markers in three chromosomes (LG 4, 19 and 22-25), the mean recombination frequency per chromosome arm was calculated as 0.989 ± 0.12 suggesting one crossover per chromosome arm. However incidences where multiple crossovers were taking place was also observed (Table S5, available in online version), suggesting that interference is not complete. The high marker density in this study probably helped to detect these events.

It was not possible to construct a genetic linkage map directly from the meiotic gynogenetic genotypic data in this study. It was not entirely clear if this was due to the nature of the data or the fact that linkage mapping softwares were not developed for this type of family. However, after defining linkage groups from the distribution of the markers in the sea bass genome assembly, we were able to order markers within these linkage groups using mapping softwares, and subsequent analyses suggest that this was a successful approach. We suggest that in any future similar studies, it would be better to produce a diploid biparental family as well as a meiotic gynogenetic family, then the recombination data could be overlaid onto the linkage map constructed from the biparental sibs, which should contain essentially the same set of markers. Guyomard et al. (2006) followed this approach to some extent ($n = 60+60$ in two biparental family; $n = 60$ in meiotic gynogenetic family), but did not use large numbers of markers in meiotic gynogenetics or describe any attempt to construct a linkage map from the meiotic gynogenetic data. Rather the strategy that he used was solely based on limited number of markers (not stated in the manuscript, pers.comm. R.Guyomard) so as to give a more accurate order to the markers by defining relative position of centromere in duplicated genome of rainbow trout and brown trout.

While there was high congruence between the genetic map from this study and the high density map of Palaiokostas et al. (2015b), six linkage groups showed low correlation in marker order between the linkage map of Palaiokostas et al. (2015b) and the genome assembly. This may reflect problems in accurate assembly of these chromosomes in this first draft sea bass genome. Therefore the data from the two linkage maps could be used in improving the genome assembly.

3.5 Conclusion

In an effort to define telomeric markers to aid reliable production of clonal lines by differentiating between meiotic to mitotic gynogenesis, the present study constructed a genetic linkage map from a meiotic gynogenetic family of European seabass. Markers located at the distal end of the chromosome are of interest with higher level of recombination. To our best knowledge it is also a first genetic linkage map based on draft genome of sea bass in meiotic gynogenetic family. Based on genetic linkage map order, this can be used for fine-tuning of the genome assembly for the future versions. Crossover frequency per chromosome arm observed was similar with the existing literature which was additionally supported by marker-centromere mapping distances therefore supports the hypothesis on high level of interference in fish species. Overall this work demonstrated the potential of next generation sequencing technologies on identifying hundreds of SNP markers in short period of time with a cost effective manner. On the basis of having identified telomeric markers, the future research will involve the verifications of such markers, which in deed should help to overcome the bottleneck of producing clonal lines via mitotic gynogenesis in European seabass. Thus, reliable production of clonal lines will robustly be achieved by detecting and eliminating any spontaneous meiotic gynogenesis from the doubled haploid isogenic G1

fish. This study provides a pilot study on the efficacy of NGS technologies for the verification of isogenic fish lines, thus should help accelerating production of such lines in fish for research related use in aquaculture.

Acknowledgments

We gratefully acknowledge support from the European Fund Aquaculture infrastructures for excellence in European fish research, AquaExcel project (FP7), and from the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland), funded by the Scottish Funding Council (grant reference HR09011) and contributing institutions. MO gratefully acknowledges the financial support of the Turkish Government, Ministry of Education for her PhD scholarship.

Chapter 4

Genome-wide verification of isogenicity of clone founders (G1) in European seabass (*Dicentrarchus labrax*) through ddRADseq

Münevver Oral^{1§}, Julie Colléter², Michaël Bekaert¹, Kerry Bartie¹, John B Taggart¹, Brendan J McAndrew¹, Marc Vandeputte^{2,3}, François Allal², Alain Vergnet², Béatrice Chatain², David J Penman¹

¹ Institute of Aquaculture, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK

² Ifremer, UMR9190 MARBEC, Chemin de Maguelone, 34250 Palavas Les Flots, France

³ GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

Author Contribution: The first draft of the present manuscript was compiled and written in full by the author of this thesis, who was also fully involved in all subsequent revisions, preparation of ddRAD library (under guidance of John Taggart), segregation analysis and subsequent genome-wide verification study and general statistics were carried out by the candidate. JC and AV produced mitotic gynogenetic fish. KB performed DNA extractions. The other co-authors contributed towards the experimental design, the analysis of sequenced reads, and machine learning algorithms.

Abstract

Isogenic clonal lines of fish are a valuable tool for aquaculture-related research, as inbred animals have been in biomedical research, yet to date they are available in only a few species. Although the production of such lines can be achieved in two generations through induced parthenogenesis (either mitotic gynogenesis or androgenesis), challenges such as potential contribution from irradiated gametes, reduced survival of doubled haploid clone founders and spontaneous, partially heterozygous meiotic gynogenetics (due to non-targeted retention of the second polar body in the mitotic gynogenesis process) hamper the successful establishment of such lines. Until recently only small numbers of genetic markers were available for genotyping and thus verification of such lines. Reliable and efficient marker technologies are needed for genome-wide screening during development of isogenic lines, and high-throughput sequencing (HTS) offers this potential. In the present study, we analysed DNA from 18 putative mitotic gynogenetics (clone founders) of European seabass that was initially genotyped and selected based on isogeny of 12 microsatellite loci, using double-digestion restriction-site associated DNA sequencing (ddRAD-seq). A total of over 31 million raw sequence reads were produced and assembled into an average of 6,830 unique ddRAD loci. Based on an average of 1,950 polymorphic single nucleotide polymorphism (SNP) loci, 17 out of 18 fish were identified as isogenic (mitotic gynogenetics), while one fish represented a clear case of spontaneous meiotic gynogenesis, with no sire contribution but heterozygous for 49% of informative maternal loci. Although these fish were genotyped first using 12 microsatellite loci which suggested isogenicity of all samples, the single meiotic gynogenetic was only detected with the higher power of genome-wide screening in one fish, proving the efficacy of HTS for this

purpose. Provided that the clonal founders are fertile, they will be used for producing isogenic clonal lines in European seabass in the next generation. Successful establishment of such lines in species of prime commercial interest in Europe is one of the objectives of the AQUAEXCEL²⁰²⁰ as a resource for aquaculture-related research.

Keywords: *Isogenic clonal lines, mitotic gynogenetics (G1), ddRADseq, European seabass, aquaculture*

4.1 Introduction

European seabass is an important farmed marine species in the Mediterranean basin. Since the 1980s, commercial scale hatchery production of seabass became profitable, and farmed production reached 152,000 tonnes in 2014, 95% of the total market for this species (FAO, 2014). European seabass and gilthead sea bream constitute the first non-salmonid marine species of commercial interest in Europe. Regardless of intensive aquaculture activities, selective breeding programmes are still in early stages (Vandeputte et al., 2009).

Selective breeding is one way to boost productivity in all farmed species, thus is widely practised and successfully applied mainly in terrestrial animals and to some extent in aquatic species. Classical breeding has improved the development of genetically improved, high yielding seed stocks particularly for relatively simple traits. However such productivity traits are, mostly, ruled by many quantitative trait loci (QTL) and their interactions with the environment thus the selection towards improved productivity is challenging. This is where clonal lines come in: such lines are a unique tool to *fix* genotypes affecting traits of interest and hence help in

dissection of QTL components of traits under investigation. Isogenic clonal individuals are more likely to express *extreme* genotypes due to additive genetic variance component between families (Bongers et al 1997a), while genetic variance within families equals to zero. Mapping panels produced by crossing such extreme genotypes create a strong QTL mapping structure to observe the segregation of traits of interest. In some cases additional complications might arise. For example in the case of studying polygenic fillet quality traits (e.g: omega-3 content, fillet colour, harvest weight) due to assessment of the phenotype can only be carried out post harvest makes selection for those traits more difficult since such fish cannot be used as selection candidate (Gjerdem 1997,2005; Johnstone, 1999; Hamzah et al., 2016).

However production of a single clonal line is not a realistic approach, simply because each clonal line represents genomic sampling of a single sperm or unfertilised egg derived from a single parental fish (androgenesis and gynogenesis, respectively). Therefore it is of great interest to propagate isogenic clonal lines from several outbred and/or domesticated stocks so as to reveal genetic variation to be compared in future studies. To this end, Quillet (1994) recommended producing as diverse as possible mitotic gynogenetics from various outbred populations, even with very few numbers of survivals in each putative mitotic gynogenetic fish so that the selection of lines could be more effectively applied with many extreme genotypes to be characterised. Similarly Bongers et al. (1997b) utilised additive genetic variance of homozygous gynogenetic families to produce and subsequently select genetically diverse early and late maturing fish with high egg quality in Common carp. This approach of establishing many DH fish in the first generation of isogenic clonal fish lines is well recognised as a better representative of specie

as whole (Robison & Thorgaard, 2011).

Although gynogenesis has successfully been applied in European seabass (Peruzzi & Chatain, 2000; Francescon et al., 2004; Bertotto et al., 2005, Colléter 2015), there are some technical limitations that can impact the effectiveness of the resultant progeny. The first limitation can be encountered as a form of potential genetic contribution from irradiated gametes. Such fragments are derived by un-optimal irradiation of gametes. As a typical trend; relatively low irradiation doses gives rise to persistent chromosomal fragments while increasing doses of irradiation can lead into motility losses in irradiated spermatozoon. As a result, the existence of chromosomal fragments from irradiated genome has been observed in numerous species in both inductions of gynogenesis and androgenesis (Arai et al., 1992; Quillet 1994; Bertotto et al., 2005; Colléter et al., 2014). Earlier studies utilised recessive morphological characters (e.g: pigmented gene of wild type Nile tilapia) as evidence of uniparental inheritance. Although such phenotypic traits have advantages of being easily observable, yet these cannot be used alone as a measure of complete inactivation of genome from irradiated parent (Pandiran and Kirankumar, 2003). Later on, marker technologies have been updated to allozymes, DNA fingerprinting, AFLPs and microsatellites as reviewed by Komen & Thorgaard (2007). The second limitation can be observed with untargeted occurrence of meiotic gynogenetics in mitotic gynogenetic group. Two interpretations were suggested to explain spontaneous occurrence of meiotic gynogenetics among the mitotic gynogenetic group by different researchers: inhibition of the first or second meiotic division or non-disjunction of chromosome pairs during these developmental processes, by Quillet et al. (1991) and (Komen et al., 1991), respectively. Alternatively, at a slow developmental rate of certain eggs

(gynogenesis protocols are severely affected by egg quality, also termed as *late maturation* effect), a ‘late’ shock might inhibit the second meiotic division instead of the first mitotic division, resulting in induced meiogynogens (Galbusera et al., 2000; Bertotto et al., 2005). The possibilities of observing spontaneous and induced meiogynogens are rather small yet given the rarity of establishing successfully mitotic gynogenetics (due to very high mortality) the effect of residual heterozygosity might be of significant as suggested by Quillet et al. (1991). Therefore verification step is needed in each generation to validate the production protocols and/or to identify isogeny comprehensively.

Identifying markers that are capable of discriminating mitotic and meiotic gynogenetics can be a challenging task. Telomeric markers will be heterozygous in meiotic gynogenetics and homozygous in mitotic gynogenetics, while centromeric markers will be homozygous in both types and thus lack discriminatory power (see Chapter 1). Important aspects to take in to account are to use (i) larger numbers of marker panels and (ii) consider the ability of markers to discriminate between mitotic and meiotic gynogenetics (Thorgaard 1983; Danzmann & Gharbi 2001).

Until recently small numbers of microsatellite markers have been used to discriminate partially heterozygous meiotic gynogenetics from 100% homozygous mitotic gynogenetics. Galbusera et al. (2000) used 5 polymorphic microsatellite loci to verify isogenicity of clone founders in African catfish (*Clarias gariepinus*), Bertotto et al. (2005) used 6 microsatellite loci in European seabass (*Dicentrarchus labrax*) while recently Alsaqufi et al. (2012) used 10 microsatellite markers in ornamental varieties of domesticated koi carp (*Cyprinus carpio*) for the same purpose. Given that many teleosts have around 22-26 chromosome pairs, including European seabass ($n = 24$), such marker sets do not even cover the karyotype of the

species of interest. This also highlights that any fragmentary paternal contribution could have been missed due to limited amount of markers used in the past.

Although microsatellites are useful as genetic markers, when used in limited numbers depending on their informative content and diagnostic power, they are likely to overlook the polymorphism under investigation. This, in conjunction with having a common allele between parental genotypes can decrease the information level of a given locus. Therefore, if small numbers are available (compared to the number of chromosome pairs in species of interest, e.g. > 25 microsatellite markers well spread along the European seabass genome can be informative, although does not cover both chromosome arms, while this number is very limited in Atlantic salmon genome assuming markers located in telomeric regions) a second round of verification is required so as to confirm the results of initial selection panel. To this end, telomeric markers offer the potential to reliably differentiate between meiotic and mitotic gynogenetics with their higher power of detecting any heterozygote contamination in the clone founder progeny (G1). Homozygosity at these loci indicates successful production of mitotic gynogenetics.

Emerging sequencing technologies have increasingly been enabling genotyping by synthesis at much lower cost and in a shorter time frame, allowing large numbers of SNP markers to be generated in one sequencing run. This provides a unique opportunity to investigate potential residual chromosome fragments that might be observed from irradiated genome and to distinguish mitotic gynogenetics from those of meiotic gynogenetics. In this study SNP markers generated by double-digest restriction associated DNA (ddRAD) sequencing (ddRAD seq; Peterson et al., 2012) were employed to comprehensively examine parental genetic contributions (particularly sire contribution) in experimentally generated putative

isogenic mitotic gynogenetic progeny. The main objectives of the study therefore were to (i) search for potential paternal contribution from UV-irradiated sperm; (ii) investigate genome-wide isogenicity in putative isogenic G1 fish in European seabass and (iii) analyse the efficacy of NGS technologies in potential false positive meiotic gynogenetics (based on initial microsatellite panel of 12 loci) so that they could be removed from the pool of mitotic gynogenetic clone founder progeny used to establish isogenic clonal fish lines in the subsequent generation.

4.2 Materials and Methods

4.2.1 Production of clone founders through mitotic gynogenetics

4.2.1.1 *Overview*

The mitotic gynogenetic European seabass families were produced at the Ifremer Experimental Aquaculture Station (Palavas-les-Flots, France), using eleven dams and eleven sires to produce twenty-two families by artificial fertilisation of the eggs with i. UV irradiated sperm (for mitotic gynogenetics) or ii. normal milt (as biparental control groups) in each dam x sire combinations. At 187 dph, all surviving fish were individually tagged and fin clips were taken. These were used for genotyping of 12 microsatellite markers and initial selection of homozygous fish was based on these loci. In total, 26 fin clips (8 parental samples, belonging to 4 families, and 18 putative isogenic clone founders, see Figure 4.1) were collected and stored in absolute EtOH and sent to the University of Stirling for the verification of isogenic status of putative isogenic clone founders (G1) in European seabass using ddRADseq.

4.2.1.2 *UV irradiation of sperm, pressure shock and husbandry*

The husbandry procedures applied to broodstock and the gamete collection were as described in Colléter et al. (2014). The UV irradiation device was composed of eight UV germicidal lamps (12 W, 254 nm, Vilber-Lourmat, Marne-la-Vallée, France) fixed above and below (four lamps each) a quartz plate which was mechanically agitated to stir sperm samples throughout irradiation. After checking for sperm motility in each sire, 0.5 ml of diluted sperm from a single male (diluted 1:20, v/v in artificial extender Seabass Gamete Short term Storage – (SGSS) Storefish (IMV Technologies, France) supplemented with pyruvate and glutamine at 0.6 and 3 mg.ml⁻¹ respectively: C. Fauvel, personal communication) was poured into an 8.5 cm diameter quartz Petri dishes (SARL NH Verre, Puechabon, France). The UV lamps were switched on at least 30 minutes before administering of irradiation dose. Optimal UV dose was checked both at the beginning and at the end of each experiment using a VLX-3W UV radiometer (Vilber-Lormat), checking both upper and the lower sources. The total UV irradiation dose applied was 320 mJ.cm⁻² based on previously optimised protocol by Peruzzi & Chatain (2000).

Artificial fertilisation was performed just after UV irradiation of sperm by adding 5 ml of (1:20) SGSS diluted sperm to 125 ml eggs (untreated, good quality) then the same volume of (125 ml) seawater (14°C, 35‰). Timing for the pressure shock was started as soon as the seawater was added to the eggs and irradiated sperm to initiate fertilisation. The labelled egg batches fertilised with UV irradiated sperm were stored in darkness at 14°C until the application of pressure shock to restore diploidy. A pressure shock of 8500 psi for 4 minutes duration was applied, at timings calculated using the equation of Francescon et al. (2004) based on the first cleavage timing which varied from 99 to 109 minutes after fertilisation. Biparental

control groups received the same procedure using normal sperm and without pressure shock (ordinary fertilisation). All experiments were performed in total darkness until the end of incubation period in a temperature controlled room maintained at 14°C.

Control and treated eggs were incubated separately in individual 40 L tanks in a dedicated recirculated water system (temperature of 14-14.5°C and salinity of 35-36‰). Embryo development was checked under a dissecting microscope (M3C, Wild Heerbrugg, Switzerland) by collecting samples from each incubator at 2-4 HAF to assess fertilisation rate at 4-8 celled stage and 50 HAF to assess further embryonic development in European seabass. Right after assessment of embryonic development specimens were returned to their incubator. Approximately a day before hatching, at 74 HAF, surviving embryos were transferred to larval rearing tanks where common garden protocol was applied to triplicated samples (6 batches of eggs consisting 3 mitotic gynogenetics groups and 3 bi-parental control groups). Larval rearing was performed in 0.5 m³ tanks in a recirculated system where water renewal between 10-20%h⁻¹ with a constant salinity of 25‰ and an oxygenated air flow of 100-120ml.min⁻¹. Larvae were kept in the dark until 12 days post hatching (DPH), corresponding to 160°C x day, when artificial lighting of 100 lux for 12 hours a day was introduced. In between 5-12 DPH, tanks were equipped with a home-made surface cleaner system to remove oil and floating debris from the water surface to ensure good swim bladder inflation during development of larva. Feeding was initiated at 12 DPH with freshly hatched naupli of *Artemia salina* supplied daily. Ordinary husbandry procedures were applied from first feeding to all groups.

Fish were individually tagged with Passive Integrated Transponder (PIT-tag) glass tags at 187 DPH, fin clipped and numbers per tank were equalised to 250 fish per tank. Survivors from experimental groups (putative mitotic gynogenetics) were genotyped at 12 microsatellite loci that were validated for gene-centromere distances previously: any heterozygosity detected in such fish indicated that they were not mitotic gynogenetics as expected, therefore they were removed from the putative isogenic G1 fish.

The determination of gene-centromere distances were carried out in 3 meiotic gynogenetic families (female A, B and C) produced by applying an early pressure shock (6 minutes AF for the duration of 2 minutes at 8500 psi) following the previously optimised protocol of Peruzzi & Chatain (2000). Based on the results of 12 microsatellite loci, 26 fin samples (8 parental and 18 putative isogenic clone founders) were provided for the detailed analysis of genome-wide isogenicity.

4.2.2 ddRAD library preparation and sequencing by synthesis

A universal salt buffer (Aljanabi & Martinez, 1997) method including SSTNE-SDS as explained by Taslima et al (2015) was used for high quality genomic DNA extraction. The concentration and the purity of each sample were assessed initially by using spectrophotometry (Nanodrop) technique and the molecular weight of DNA was assessed by agarose gel electrophoresis by observing intact high molecular weight DNA bands. The final DNA concentration of each sample was carried out using a fluorescent assay, Qubit dsDNA BR Assay Kit (Invitrogen, UK), which only quantifies double stranded DNA molecules prior to ddRAD library construction. Each sample was diluted to a concentration of 5 ng/ μ L in 5 mM Tris, pH 8.5 based on Qubit reading. Table S1 (Appendix) gives detailed information of

the samples used in the present study.

The ddRAD library preparation protocol followed essentially the methodology originally described in Peterson et al. (2012) with slight modifications explained by Palaiokostas et al. (2015b). The in-house procedure modified here differed from the original protocol in one key matter: pooling was applied at the earliest stage (right after barcoding) and pooled samples were processed within a single tube rather than processing each sample singly throughout. Given the limited number of samples (twenty-six fish in total: eighteen putative clone founders and eight parents, Figure 4.1) used for the verification study replications were used so that higher coverage per individual sample could be achieved. Parents were triplicated while putative clonal fish were duplicated in the library (Table S1, available in electronic version).

Each sample (0.015 µg DNA) was digested at 37°C for 90 minutes with *SbfI* (rare cutter, recognising the CCTGCA|GG motif) and *SphI* (common cutter, recognising GCATG|C motif) high fidelity restriction enzymes (New England Biolabs; NEB), using 20U each enzyme per microgram of genomic DNA in 1× CutSmart Buffer (NEB). No heat inactivation was performed at any stage of the procedure. The restriction digestion reaction volume per individual sample was 6 µL (3 µL of 5ng/µL gDNA + 3 µL of RE MMix). Individual-specific combinations of P1 and P2 adapters (Table S1, available in electronic version), each with either a unique 5 bp or 7 bp barcode, were ligated to the RE fragmented DNA at 17 °C for the first hour then at 22 °C for two more hours (3hrs in total) by adding 0.6 µL 100 nmol/L adapters, 0.15 µL 100 mmol/L rATP (Promega), 0.25 µL 10× CutSmart Buffer (NEB), 0.12 µL T4 ligase (NEB, 2 M U/mL) and reaction volumes made up to 12 µL (3 µL *SbfI*:*SphI* barcode mix 1:10 v/v + 3 µL of ligation MMix) with nuclease-

free water for each sample. The barcodes were selected to differ from each other by at least 3 bases. Following ligation reactions all samples were combined in a single pool (for one sequencing lane) and purified by MinElute PCR clean up kit (Qiagen, Invitrogen).

Size selection (320-590 bp) was performed by agarose gel separation and was followed by gel purification and PCR amplification. A total of 50 μ L of the amplified library (12 cycles) was purified using an equal volume of AMPure beads. After eluting into 20 μ L EB buffer (MinElute Gel Purification Kit, Qiagen), the library was quantified before and after bulk PCR amplification by Qubit (dsDNA HS, Invitrogen). The ddRAD library was then diluted down to 2.5 nM final library concentration by using freshly prepared 0.2M NaOH / 1% Tween 20 (Alpha Laboratories). Denaturation of final library was achieved by using both chemical (NaOH) and heat treatment (2 minutes incubation at 98°C, then chilled on ice for 5 minutes) according to Illumina's protocol. The final library was mixed with 4% PhiX (control library of Illumina) to reach desired loading concentration of 10.6 pM of the library. This was loaded on to MiSeq cartridge belonging V2-300 kit for paired end sequencing.

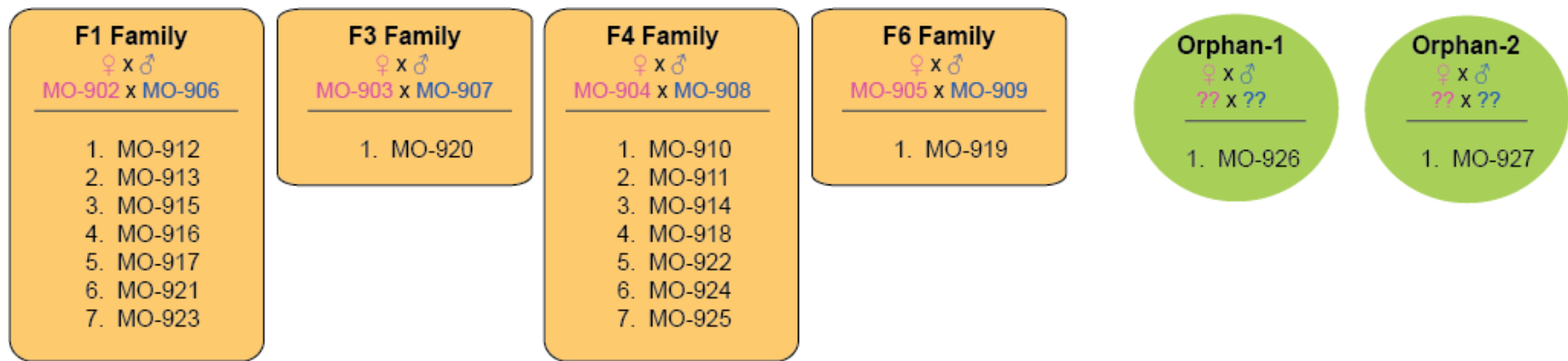


Figure 4.1: The pedigree of the samples.

4.3 Data Analysis

4.3.1 Sequence Quality Control (QC)

A quality check of raw data was initially assessed by the in-built software of the sequencer, MSC (Miseq, Illumina). Later on FastQC v.0.11.3 (Andrews, 2010) was used to generate a comprehensive quality report. Reads of low quality (Phred score under 30), missing the restriction site or with ambiguous barcodes were discarded by using *process_radtags* module implemented in Stacks (Catchen et al., 2011). This module also demultiplexes data after inspecting barcodes and ddRAD cut-sites are intact. The filtered read files were then renamed to reflect sample names for ease of further analysis and the barcodes removed. Retained reads were trimmed to a length of 140bp. This trim was not essential given that the quality of per base sequence in both P1 and P2 reads were falling into *very good quality calls* (Phred>30), yet trimming was still applied considering that the end of the sequencing run in each fragment is more likely to represent sequencing errors than any other parts (see Report S1 and Report S2 where lower whiskers in per base sequence quality plot fall to reasonable quality read area as the sequencing continues, note the decreasing quality trend of the sequencing run which is a typical phenomenon, available online).

4.3.2 SNP calling

The trimmed reads (140 bp) were sorted into loci and genotyped using the Stacks pipeline v1.40 (Catchen et al., 2011). The likelihood-based SNP-calling algorithm (Hohenlohe et al., 2011) implemented in Stacks evaluates each nucleotide position in every ddRAD-tag of all individuals, thereby differentiating true SNPs from sequencing errors. Reads were aligned to the reference genome assembly of seabass (dicLab_v1

accessed on May 2016) using Bowtie 2 (Langmead & Salzberg, 2013) and the output of the program in the form of SAM files were fed into Stacks pipeline *rep_map.pl* assembly module for SNP calling. Minimum identical number required to create a stacks (-m) of 6 was used. Each family was analysed in separate batches in the same catalogue. In order to compare the isogenicity of MO-926 and MO-927 all samples were analysed in population setup so that lack of parental information in these samples would be eliminated (see Figure 4.1).

Once SNP calling was completed the following filters were applied prior to extracting genotypes: ddRAD loci shared among 70% of all the samples with both parental genotypes available, carrying up to 3 SNPs and 4 alleles.

4.3.3 Investigation of putative sire contributor loci

Initial examination of heterozygotes was carried out on the web interface of Stacks. The aforementioned filters were applied to each family with the different parental genotype combinations. The ones where dam and sire had distinctive alleles (eg: aa/bb dam and sire respectively) were particularly chosen to examine any contribution from irradiated sire genome in putative clone founder progeny. These genotypes were also useful to observe any potential sire contribution to progeny. Following initial examination, the main analysis was carried out on the genotype files extracted with the same filter on Excel. Each family was individually checked for the segregation of each parental genotype by simply applying data filters. This analysis involved counting missing genotypes, and dam & sire informative genotypes individually and summarising the outcome for each offspring. In the case of observing an unexpected genotype (heterozygotes) in putative clone founders, each locus was cross-checked on the web interface for the verification of a given valid locus. A valid locus was defined based on

high coverage (>10) for each allele and the nature of the ddRAD reads. For example some repetitive reads are more likely to possess sequencing errors: such loci were removed from the dataset.

4.4 Results

4.4.1 ddRAD sequencing

Sequencing of ddRAD loci was carried out on 26 individuals. A total of 31,113,626 reads (each 162 bp long) were obtained at the end of one sequencing run (Figure 4.2). Following process_radtag module of Stacks pipeline low quality reads (Phred33, quality score under 30) (469,458), ambiguous barcodes (3,411,708) and ambiguous RAD tags (128,295) were removed. This subsequently resulted in 87% of the raw reads being retained (27,104,165). Filtered reads were assembled into an average of 6,830 RAD loci per individual. The raw sequence data from this study were deposited at the EBI Sequence Read Archive (SRA) with the accession number PRJEB15131.

Table 4.1 shows the sequencing statistics regarding to overall alignment rate of samples against the reference genome assembly of *D. labrax* (dicLab_v1) and average coverage per locus achieved per sample at the end of one sequencing run. High alignments rate were achieved with an average of above 100x coverage per sample. Thus reliable and robust sequence data was produced in the present study.

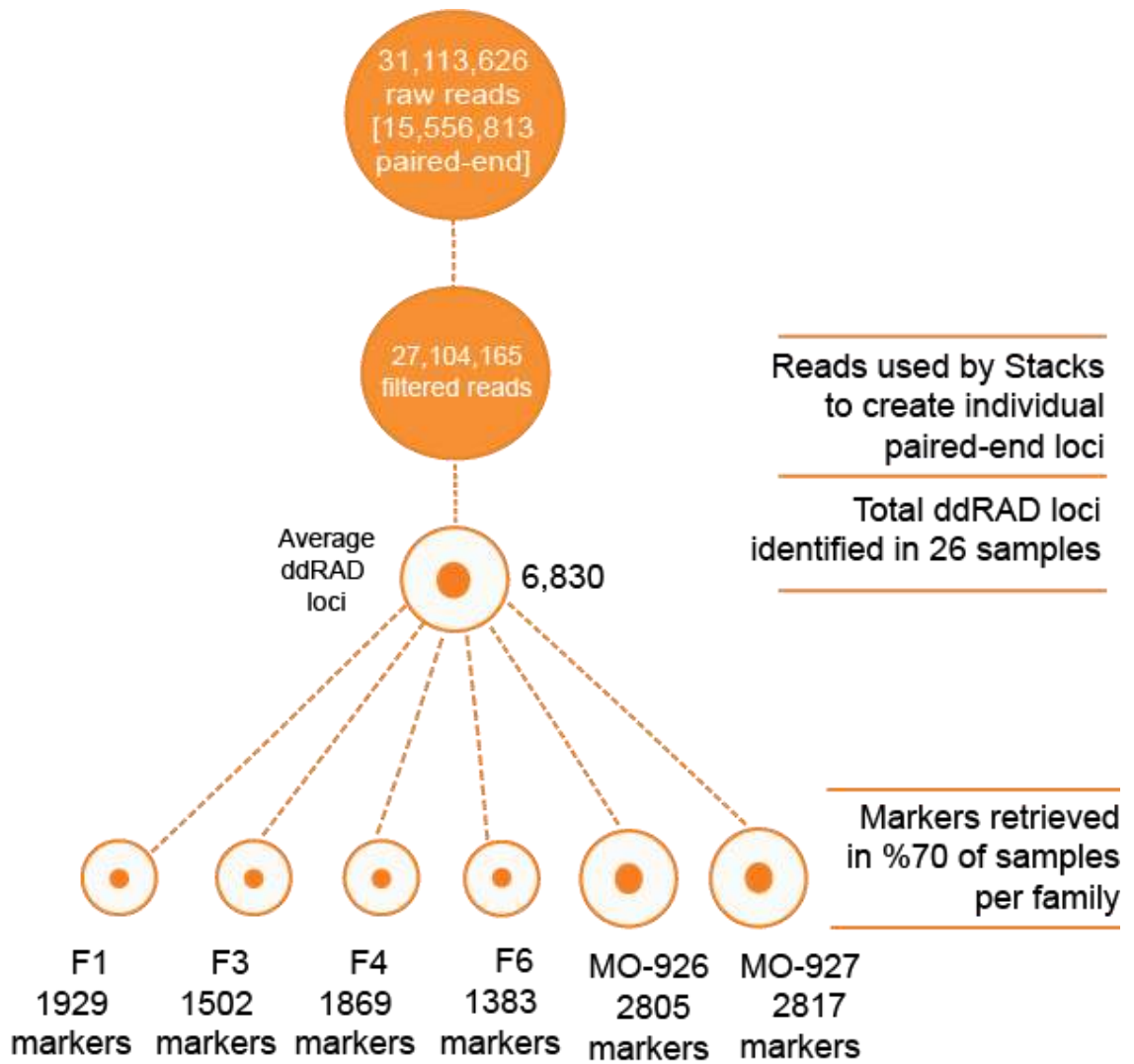


Figure 4.2: Sequencing and ddRAD-tag summary. Details of the number of reads before and after filters (orange disk) followed by the reconstructed number of ddRAD loci after filtering. The final number represents the polymorphic loci (markers) available per family after removing missing genotypes.

Table 4.1: Summary of overall alignment rate of samples against the reference genome assembly of *D. labrax* (dicLab_v1) and overall depth of coverage achieved per sample at the end of one sequencing run.

Alignment rate to <i>D. labrax</i> genome assembly	Average coverage per locus
MO-902.log:98.06%	140.715x
MO-903.log:98.12%	78.536x
MO-904.log:97.94%	155.235x
MO-905.log:98.15%	102.067x
MO-906.log:98.10%	195.418x
MO-907.log:98.07%	165.412x
MO-908.log:97.91%	148.602x
MO-909.log:98.02%	151.843x
MO-910.log:97.77%	146.16x
MO-911.log:98.06%	106.378x
MO-912.log:98.14%	158.602x
MO-913.log:97.95%	140.594x
MO-914.log:97.81%	148.729x
MO-915.log:98.04%	94.0837x
MO-916.log:97.93%	147.294x
MO-917.log:98.09%	112.392x
MO-918.log:97.68%	183.664x
MO-919.log:98.16%	121.944x
MO-920.log:98.07%	135.036x
MO-921.log:98.00%	126.041x
MO-922.log:97.76%	127.511x
MO-923.log:98.09%	153.958x
MO-924.log:97.67%	121.394x
MO-925.log:97.95%	172.138x
MO-926.log:97.95%	121.029x
MO-927.log:97.80%	126.795x

4.4.2 Distribution of ddRAD alleles

Overall, high numbers of polymorphic ddRAD loci were identified in each family, ranging from 1383 to 2817. Families possessing only one progeny, F3 and F6, resulted in lower number of loci (1502 and 1383 loci respectively) compared to families with 7 progeny, F1 and F4 (1929 and 1869 loci respectively) (see Figure 4.2). The average frequency of heterozygous loci ranged between 0.21 to 1.75 % with an exception of 31% heterozygous loci detected in the only progeny of F6 family, confirming that the progeny is not isogenic.

Limited proportions of heterozygotes (<2% in most families) detected among the other putative isogenic clone founders were reflected in increased frequencies of homozygote genotypes, ranging between 96.93 and 99.46% among all families, except from the F6 progeny. Details of the distribution of ddRAD genotypes per family are given in Tables 4.2 and 4.3. The sequences of the markers used for the verification of isogenic clone founders in each family are provided in dataset S1 (available in electronic version).

4.4.3 Investigation of putative sire contributor loci

Most of the few heterozygous genotypes observed among the putative clone founders fell into the category of ignorable genotypes, where the locus had an allele depth of less than 10 reads, or had escaped from the initial filtering (for any of the alleles scored and/or locus had more than 3 alleles in a diploid organism which was a clear indication of an error.) A small proportion of annotated genotypes detected were the end results of automated default corrections of the pipeline (see Discussion) and were ignored within the scope of this study. There were cases where a limited proportion of detected heterozygotes in the putative

isogenic clone founders were due to small scale duplications events taking place in the genome of *D.labrax*. Such cases were easy to detect with extremely high allele depths, over 400 in most cases, and shared genotype where both parental and progeny represented the same genotype as *abxab* crosses. There were limited cases where the dam was misgenotyped. Detailed investigation of such loci revealed that Stacks ignored a second allele due to its small proportion. For example, a given locus genotyped as *aa/bb* (allele depth of 148/260) in dam and sire respectively is expected to give rise to *aa* (an average of 100 reads) progeny only. However in such cases progeny represented *aa* and *bb* genotype indicating that dam's genotype should have been *ab*, detailed investigation of individual Stacks used for genotype calling in dam confirmed that the dam was misgenotyped in the given locus (in total 7 loci among all families, see Table 4.4 and 4.5) due to the low number of reads detected for the alternate allele which was automatically ignored by the pipeline (e.g: 17 reads for alternate allele [*a* in this case] and 131 reads for the other allele [*b*): one of the limitation of the Stacks pipeline is that genotype calls are in favour of homozygotes by the nature of the pipeline, see Chapter 7 for detailed discussion). Such cases were detected after manually checking each of these loci. Overall among the heterozygote genotypes detected, which were less than 2% on average in putative clone founders in European seabass, there were no clear signs of sire contribution to progeny in any of the families investigated, including F6 with over 30% heterozygotes detected, as well as the two "orphans". Details of the investigation of putative sire contribution per family are given in Tables 4.4 and 4.5.

Table 4.2: The distribution of ddRAD alleles in F1 and F3 families

Genotypes	F1 family							F3 family
	MO-912	MO-913	MO-915	MO-916	MO-917	MO-921	MO-923	MO-920
Individuals								
*Missing genotypes	46	43	25	58	29	49	45	40
Heterozygotes (%)	11 (0.58)	11 (0.58)	9 (0.47)	9 (0.47)	12 (0.62)	10 (0.52)	16 (0.84)	26 (1.75)
Homozygotes	1900	1903	1923	1890	1916	1898	1896	1456
Polymorphic ddRAD loci	1957	1957	1957	1957	1957	1957	1957	1522

*: Missing genotypes can arise due to scoring of loci in parents but not in the progeny.

Table 4.3: The distribution of ddRAD alleles in F4, F6 families and two orphans (MO-926 & 927)

Genotypes	F4 family							F6 family	Orphans	
	MO-910	MO-911	MO-914	MO-918	MO-922	MO-924	MO-925	MO-919	MO-926	MO-927
Individuals										
*Missing genotypes	35	48	48	60	44	33	61	41	131	116
Heterozygotes (%)	11 (0.59)	14 (0.76)	16 (0.87)	14 (0.77)	12 (0.65)	10 (0.54)	16 (0.88)	422 (31.07)	6 (0.21)	10 (0.35)
Homozygotes	1823	1807	1805	1795	1813	1826	1792	936	2790	2801
Polymorphic ddRAD loci	1869	1869	1869	1869	1869	1869	1869	1399	2927	2927

*: Missing genotypes can arise due to scoring of loci in parents but not in the progeny.

Table 4.4: Summary of putative sire contributor loci in F1 and F3 families

Genotypes	F1 family							F3 family
	MO-912	MO-913	MO-915	MO-916	MO-917	MO-921	MO-923	MO-920
Individuals								
[‡]Hets detected (%)	11 (0.58)	11 (0.58)	9 (0.47)	9 (0.47)	12 (0.62)	10 (0.52)	16 (0.84)	26 (1.75)
*Ignorable loci	6	7	4	3	7	5	8	6
Annotated loci	-	1	-	3	1	-	2	12
[#]PSSD loci	5	3	5	3	3	5	6	5
Misgenotyped dam	-	-	-	-	1	-	-	3
Polymorphic ddRAD loci	1957	1957	1957	1957	1957	1957	1957	1522

[‡]: Total heterozygote genotypes detected with their percentage

[#]: Potential small scale duplicated loci

*: Loci that are justifiable to ignore with;

-Allele depth < 10 reads per loci

-Locus possess >3 alleles

Table 4.5: Summary of putative sire contributor loci in F4, F6 families and two orphans (MO-926 & 927)

Genotypes	F4 family							F6 family	Orphans	
	MO-910	MO-911	MO-914	MO-918	MO-922	MO-924	MO-925	MO-919	MO-926	MO-927
Individuals										
[‡]Hets detected (%)	11 (0.59)	14 (0.76)	16 (0.87)	14 (0.77)	12 (0.65)	10 (0.54)	16 (0.88)	422 (31.07)	6 (0.21)	10 (0.35)
*Ignorable loci	5	6	7	6	5	5	7	See Table 4.6	1	3
Annotated loci	2	2	5	1	3	-	2		-	1
[#]PSSD loci	4	5	4	6	4	5	6		5	6
Misgenotyped dam	-	1	-	1	-	-	1		-	-
Polymorphic ddRAD loci	1869	1869	1869	1869	1869	1869	1869	1399	2927	2927

4.4.3.1 *Meiotic gynogenetic detected in F6 family*

A high level of heterozygosity (31.07%) was detected in the single progeny in this family. None of the heterozygote genotypes observed in this fish showed any sign of sire contribution. Besides, the rest of the progeny, similar to other families, were carrying a powerful signature of successfully applied sperm UV irradiation in gynogenetics. Thus the only progeny MO-919 was an end result of spontaneous occurrence of meiotic gynogenesis, as opposed to mitotic gynogenesis as a result of recombination within female genome.

Once the assumption that the sire does not contribute to the progeny can be made based on the above analysis, the focus was then shifted towards heterozygous female markers (*ab* cross, 824 polymorphic loci). These were individually investigated to observe the distribution of homozygotes versus heterozygotes along each LG/chromosome (see Table 4.6). Out of 824 heterozygous female markers, 405 loci were heterozygote and represented 49% of the informative maternal loci. There was no sign of sire contribution in the maternal informative loci either.

Table 4.6: The distribution of female heterogametic markers in the F6 family

Position	Total loci	Heterozygotes	Homozygotes
LG10	43	1	42
LG11	37	33	4
LG12	37	37	0
LG13	10	7	3
LG14	33	0	32*
LG15	66	64	1*
LG16	27	0	27
LG17	27	13	14
LG18-21	21	1	20
LG19	29	7	22
LG20	47	4	42*
LG22-25	54	43	11
LG24	27	16	11
LG1A	15	2	13
LG1B	34	34	0
LG2	30	2	28
LB3	20	0	20
LG4	31	27	4
LG5	2	2	0
LG6	21	17	4
LG7	31	20	11
LG8	37	4	33
LG9	33	29	4
LG X	34	0	34
UNK	79	42	37
	824	405	419

*: One missing genotype detected in indicated LGs was due to scoring of loci in the parents but not in the progeny.

4.5 Discussion

In the present study we used the genomic DNA of 18 putative clone founders (initially selected based on 12 microsatellite markers) with their parents, to produce a reduced representation library. This study was the part of a project, which aimed to produce high numbers of clone founders to establish isogenic clonal lines in European seabass as part of the AquaExcel (EU, FP7) and AquaExcel2020 (Horizon²⁰²⁰) projects. However, high numbers of untargeted meiotic gynogenetics were produced alongside fully homozygous progeny in many families, as detected by the microsatellite panel, highlighting the need for a large number of DNA markers to distinguish mitotic gynogenetic individuals more reliably (Colléter 2015). See Table 4.7 summarises the results of microsatellite and the ddRADseq analysis.

Table 4.7: Summary table of the number of putative mitotic gynogenetics produced and genotyped using a panel of 12 microsatellite markers initially, followed by screening using ddRADseq of individuals homozygous for the microsatellite panel in European seabass.

Family	First screening (based on 12 microsats)		Second screening (based on ddRADseq)		Results ⁵
	Survivors of putative mitotic gynogenetic group ¹	Mitotic gynogenetics identified (%) ²	Mitotic gynogenetics analysed ³	Mitotic gynogenetics identified ⁴	
F1	650	12 (1.84)	7	7	All mitotic gyno
F3	7	1 (14.28)	1	1	Mitotic gyno
F4	18	16 (88.88)	7	7	All mitotic gyno
F6	19	1 (5.26)	1	0	Meiotic gynogenetic
MO-926	n/a*	n/a*	1	1	Mitotic gyno
MO-927	n/a*	n/a*	1	1	Mitotic gyno
Total	694	30	18	17	

¹: Total number of fish survived in experimental mitotic gynogenetic group

²: Mitotic gynogenetics identified based upon 12 microsatellite locus, the percentage of success rate is given in parenthesis based on the same marker technology

³: Total number of putative mitotic gynogenetics (based on microsatellite data) analysed by ddRADseq. In total, 18 mitotic gynogenetics were send for further ddRADseq analysis due to reduced survival since initial microsatellite genotyping.

⁴: Total number of mitotic gynogenetics identified based on ddRADseq

⁵: The end result of ddRADseq analysis per family

*: Parental information was not available

Two clear conclusions can be drawn from this table: (i) the survival rate of mitotic gynogenetics are quite low and (ii) there is a clear female effect as observed in the progeny of F4 family where 88% (16 out of 18 progeny) were initially detected as mitotic gynogenetics. This was later confirmed with the results of ddRADseq data indicating 100% success rates in F4 family based on ddRADseq data.

Until recently, only a relatively small number of genetic markers were available or used for the verification of isogenic clonal fish lines. This presented two main limitations when trying to discriminate between meiotic and mitotic gynogenetic offspring within a family. The limited number of loci used (less than 10 in most studies) would not be enough to ensure even a minimum of one marker in all linkage groups. This increases the possibility of missing residual fragmentary paternal contribution that might occur if the UV treatment of the milt was suboptimal. The second issue is that on average meiotic gynogenetic offspring are 50% homozygous and that a locus located closer to the centromere is less likely to go through a crossover event, thus its diagnostic power is lower, compared to a locus located on the telomeric parts of the chromosome where a crossover is more likely to happen. Many studies have not validated the position of their markers and their relative recombination rate. The literature has examples of both cases. Lahrech et al. (2007) recognised the higher diagnostic power of markers at the distal end of chromosome arms and used 34 polymorphic loci of which 27 were validated in four meiotic gynogenetic families and 8 telomeric markers were used as true diagnostic markers to confirm complete homozygosity in barfin flounder (*Verasper moseri*). Khan et al (2014) used 87 microsatellite loci to verify fully inbred females of a Nile tilapia (*Oreochromis niloticus*) clonal line (previously developed by gynogenesis) using microsatellite DNA markers without validating

the recombination frequencies of the loci used for genotyping, thus assuming a high number of genetic markers provides an accurate verification baseline.

The samples used in the present study were first genotyped at 12 microsatellite loci that had previously been validated on three meiotic gynogenetic families with a total number of 96 progeny for their gene-centromere distances. Following identification of recombination frequencies of each microsatellite loci a large-scale genotyping was carried out on 20 parents, 831 putative mitotic gynogenetic progeny (experimental group) and 831 bi-parental control progeny before selection of completely homozygous putative mitotic gynogenetics (Colléter, 2015).

Using a small number of markers increases the risk of obtaining a false positive where the spontaneous occurrence of a meiotic gynogenetic is falsely concluded as being a mitotic gynogenetic based on homozygosity of a few markers. This can be a result of using microsatellite loci with lower recombination frequency (such regions are homozygous in both meiotic and mitotic gynogenetics therefore not informative) and/or simply there are too few markers. To address these issues here we have utilised the high-throughput power of ddRADseq platform to inspect the segregation of alleles in thousands of loci that are well-spread throughout the genome of European seabass.

Since each family was produced by applying mitotic gynogenesis, the resultant progeny were expected to be 100% homozygous and inbred. Hence it was expected that the vast majority of the ddRAD loci would be homozygous with only a few loci that carry probable small scale duplications and appear as heterozygotes. A small portion of loci (less than 5% is acceptable) that do not follow traditional Mendelian segregation is commonly observed in any high-throughput sequencing data analysis. There are many potential reasons for observing such loci some of

which are sequencing errors, assembly related issues or small scale duplication events taking place in the genome which might be essential for the survival of the rare mitotic gynogenetics (Gu et al., 2002). Although sequencing errors are filtered to ensure high quality reads (Phred score >30) there is a possibility of carrying a limited amount of incorrect base calls, particularly at the end of reads as the sequencing by synthesis continues. As part of an assembly procedure the reads were first aligned to the reference genome assembly of European seabass (*dicLab_v1*). In the case of any imperfections or incorrect base calls in the reference genome assembly or mutation(s) these cannot be dealt with properly by the pipeline and are processed as variants. However, an average of less than 2% of odd marker frequencies was in the acceptable range thus was removed from the dataset after careful inspection of each locus for potential sire contribution (see 4.4.3 section). Limited proportions of heterozygotes (< 2% in all mitotic gynogenetic families) detected among the putative isogenic clone founders, ranged between 0.21% as minimum in orphan-1 (MO-926) and 1.75% as maximum in F3 family were also reflected to increased frequencies of homozygote genotypes: ranged between 96.93 to 99.46% among all families except from the F6 progeny. Such events which are random and rarely observed are well-accepted phenomena of biological systems where some degree of imperfections is perceived.

The markedly high numbers of total loci detected in two isogenic clone founders, MO-926 and MO-927 (see Figure 4.1), was due to the analysis method. Since MO-926 and MO-927 did not have parental DNA provided, all samples were analysed in the *population* analysis program implemented in Stacks with the same parameters of minimum identical number required to create stacks (-m) of 6 applied to both population and family based analysis module (this was applied later to each family

where parental information was available). As a general rule less filtering is applied within the population analysis module due to the lack of matching each locus to parental genotypes hence results in significantly higher number of loci compared to family based analysis with more stringent constraints. Any allele that does not match to the parental genotypes is removed from the dataset in family based analysis and therefore results in fewer loci that are truly shared in both parents and the progeny (Mendelian fashion). As the number of unique stacks (an average of 6,830) identified were similar among all the samples provided solid evidence that *population* analysis programme applied less constraints to MO-926 and MO-927 thus produced more loci to screen for the genome-wide isogenicity in both clone founders.

A limited portion of heterozygotes (< 2%) shared among the isogenic clone founders were more likely results of a probable small scale duplication (PSSD) events. These loci shared a pattern of very high coverage (>300 in most cases) enforces the hypothesis of having duplicated regions while the rest of the loci had an average allele depth of around 120 reads (Table 4.1). There were also a few annotated genotypes which were the results of automated default corrections in the pipeline. These are mostly triggered in the case of having a stack with some reads having more mismatches than set-up criteria, defined as messy stacks. In these cases the pipeline tries to match parental genotype with the offspring by pushing over the limits of set-up criteria(s), thus producing annotated genotypes in capital letters so that the user can either accept these or simply remove them from the data. Although they can be of help in some cases, such as population genetics studies, for the purposes of the present research where verification studies were undertaken, annotated genotypes were ignored for the sake of removing any unreliable loci.

The only progeny of F6 family, MO-919, was initially assigned to be a mitotic gynogenetic based on 12 microsatellite markers (1 locus, *Dla0016*, was excluded from the dataset due to low recombination rate). However, only 7 loci (5 loci had *ab/aa* and 2 loci had *ab/ac* genotypes shared by dam and sire respectively) were female heterozygotes while 3 loci were female homozygotes and had the same allele shared with the sire (2 loci *aa/aa* and 1 locus *aa/ab* dam and sire respectively) in F6 family. None of the loci had a distinctive set of alleles in parents while one locus had the same alleles in both parents as *ab/ab* genotype thus was non-informative. Furthermore two of the microsatellite loci (*Dla0104* and *Dla0106*) were located on the same linkage group, LG2, as well as being female heterozygotes. As they were physically very close to one another, and both loci represented the same genotype (*ab/aa* in dam and sire respectively), this reduced the power of microsatellite genotyping in MO-919. In the ddRADseq analysis however, high level of heterozygosity (31% in total, Table 4.5) was detected in the same individual, MO-919. There was no sign of sire contribution to the progeny therefore once this was confirmed, female heterozygous markers (824 loci) were inspected of which 49% were heterozygous (405 loci) while 51% were homozygous (419 loci). The distribution of female heterozygous markers were found throughout all linkage groups and ranged from a minimum of 0.24% (with 2 loci) on LG 5 to 8.00% (with 66 loci) on LG 15 (see Table S2 where %heterozygosity was plotted against genome assembly (Mbp) in each linkage group from both the present study and the previous meiotic seabass data, chapter 3. A limited number of microsatellites were plotted on each graph where possible). Moreover, the female heterozygous microsatellites used fell into homozygosity blocks in LGs, confirming that both marker technologies (microsatellites and SNPs) were in accordance (Table

S2, available in electronic version). In the present study, using the high-throughput power of next generation sequencing technologies provided a larger number of markers, almost evenly distributed along the LGs, and increased resolution thus allowed us to detect a meiotic gynogenetic which was previously classed as a mitotic gynogenetics based on 11 verified microsatellite markers. This shows that refined molecular genetic techniques such as ddRADseq are more promising to detect greater isogenicity than less genomically comprehensive assays of a small microsatellite panels (Mesak et al., 2014).

4.5.1 Conclusions

In an effort to verify genome-wide homozygosity of putative isogenic clone founders (G1) in European seabass, the present study utilised the high resolution power of next generation sequencing technology, starting from selected lines of interest as parents to the first generation clone founders (G1) in European seabass via mitotic gynogenetics. This work clearly demonstrated that 17 out of 18 fish, initially screened with 12 microsatellites, were homozygous based on an average of 1,950 SNP markers that are well-distributed throughout the genome of *D. labrax*. However, one fish represented a clear case of a spontaneous meiotic gynogenetic, with no sign of sire contribution yet a high level of heterozygosity (49%) originating from female recombination. Although all samples analysed in the present study were previously genotyped using 12 validated microsatellite loci which suggested the homozygosity of all samples, the single meiotic gynogenetic was only detected with the higher power of genome-wide screening. This not only proves the efficacy of NGS but also clearly demonstrates that less genomically comprehensive marker technologies such as microsatellites might give rise to *false*

positive identification, when used in smaller numbers, while NGS technologies provides more stringent evidence for the verification of genome-wide homozygosity as successfully demonstrated in the present study. The two-step verification approach used in the present study provides a realistic framework where initial mass selection of the putative mitotic gynogenetics were carried out using previously validated microsatellite for their recombination frequencies in a panel to reduce the numbers, and then the putative mitotic gynogenetics were further screened to confirm the genome-wide isogenicity of doubled haploid progeny by using genomically more comprehensive analysis of ddRADseq. Therefore, future research concerning verification of isogenic clonal fish lines is encouraged to apply such approach where markers are available for initial genotyping in species of interest. Taken together, this and the previous study hold the promise of reliable establishment of isogenic clonal lines in European seabass in the successive generation providing the clone founders (G1) are fertile. This is one of the main objectives of the AQUAEXCEL²⁰²⁰ particularly in species of prime commercial interest in Europe as a resource for aquaculture-related research.

Chapter 5

Verification of isogenic nature of clonal lines in the Atlantic salmon (*Salmo salar*) through ddRADseq

Münevver Oral^{1§}, John B Taggart¹, Stefanie Wehner¹, Brendan J McAndrew¹, David J Penman¹, Per Gunnar Fjelldal² and Tom Hansen²

¹ Institute of Aquaculture, Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK

² Institute of Marine Research (IMR), Matre Research Station, NO-5984 Matredal, Norway

Author Contribution: The first draft of the present manuscript was compiled and written in full by the author of this thesis, who was also fully involved in all subsequent revisions. DNA extraction, preparation of ddRAD library (under guidance of John Taggart) and verification study was conducted step by step by the candidate. DJP carried out haploid production in Atlantic salmon while PGF & TH produced putative clonal fish. SF carried out blast search to remove multi-copy loci. The other co-authors contributed towards the experimental design and revisions of the manuscript.

Abstract

Farmed Atlantic salmon (*Salmo salar*) is the dominant cultured aquatic species in Europe by production and value. Thus genomic resources are well established compared to other teleosts (after that of model fish species such as zebrafish or medaka). However, isogenic clonal lines have not been successfully established. The main constraints include the low survival of doubled haploid clone founders (produced through androgenesis or mitotic gynogenesis) and the ability to discriminate between such doubled haploids and fish with biparental inheritance (arising through failure of gamete irradiation) and meiotic gynogenetics (arising through untargeted spontaneous retention of the second polar body in gynogenetics). In this study, we used ddRADseq DNA sequencing to analyse the development of putative isogenic clonal lines in the Atlantic salmon starting from outbred parents to putative homozygous clone founders (G1) and to the putative isogenic clonal progeny (G2). Haploid gynogenetic embryos were analysed in parallel as a control to assist the identification of duplicated loci resulting from the ancestral tetraploid nature of Salmonidae family. A total of 46 DNA samples were used as a template to generate a ddRAD library which produced over 35 million raw reads, resulting an average of over 1,230 polymorphic SNP loci, G1 progeny were heterozygous at 8.7% while G2 families were heterozygous at 22-28% loci. All polymorphic loci were Blast searched against the three available genome assemblies of salmon to remove multi-copy loci. Single copy loci (22% of total polymorphic ddRAD loci in each family) showed exclusive transmission of maternal alleles among the six isogenic clone founders in G1 family. Varying levels of sire contribution (10-25%) were detected among the G2 families. A similar analysis using microsatellites markers (18 loci previously genotyped and 9 loci from the present study) all suggested isogenicity of the both G1 and G2 families. The existence of non-maternal

(sire alleles) among all members of the five G2 families suggests sub-optimal UV irradiation during the propagation of these putative clonal families. This study shows the utility of NGS technologies to discriminate between the different offspring types generated by different ploidy manipulations. The complications associated with the accurate identification of genotypes in a species with duplicated genomes are possible depending on the quality of the genome assemblies available. Reliable establishment of isogenic clonal lines in the Atlantic salmon, prime commercial species for Europe, is one of the objectives of AQUAEXCEL²⁰²⁰ as a resource for aquaculture related research.

5.1 Introduction

The Atlantic salmon, *Salmo salar*, is a leading aquaculture species mostly in the north Atlantic and it is increasingly cultured in Chile and Tasmania (Australia) in the southern hemisphere. It is the most important cultured fish in Europe (2,326.288 tonnes, FAO 2014) as well as being the most farmed member of the Salmonidae family worldwide (Bourret et al., 2013). The high interest in Atlantic salmon is not only limited to its commercial value but also involves the scientific, social and ecological importance of the species (Crisp 2000). The species represents a premium niche product being the number one food export for countries such as Norway and Scotland. Alongside economic value, the species is also considered as an established recreational asset due to its contribution to sport fisheries (Crisp, 2000). Given the variety of interests around the Atlantic salmon scientific interest has significantly increased over the last four decades.

Within the course of evolution the Atlantic salmon genome, in common with other members of Salmonidae family has experienced a whole genome duplication (WGD) event which took place 25 to 100 million years ago (Allendorf et al., 2015; Ohno et al., 1967). This duplication event termed as autotetraploidy; occurred as a result of tetraploidisation within the same ancestral chromosome complement. A recent study by Crête-Lafrenière et al. (2012) estimated a more precise date of 59.1 mya (with a confidence interval of 63.2-58.1 mya) for the duplication age of the family. Such events have undoubtedly provided massive amounts of raw material for adaptation, innovation and survival since the entire genetic content of the organism is doubled (Opazo et al., 2013). For example, it has been suggested one of the most important outcomes of WGD is that it gave rise to novel genes for adaptation (Glasauer & Neuhauss, 2014).

Fundamental advantages aside, WGD events and their subsequent modification can complicate our understanding of such genomes. WGD gives rise to two identical copies of the entire genome in daughter genes, which are called paralogous sequences (PSVs), fixed sites with no polymorphism. These are identical to one another and functionally redundant right after duplication. Ohno et al. (1967) suggested that such redundant genes are unique successors of the new genes essential for evolutionary innovation. Detecting paralogous sequence variants immediately after duplication is an easy process. However, complications arise as the genome evolves in time by either gaining novel functions and or sub-functions or even losing functions that are no longer essential for survival via deleterious mutations. These mutations continue to accumulate until required structural features of the gene are either completely functional or lost from the genome. This selective evolutionary process makes it very complex to detect fixed sites as variations will be introduced into paralogous sequences, which are termed as multi-site sequence variants (MSVs). In addition to PSVs and MSVs, another type of sequence variant, called single nucleotide polymorphisms (SNPs) are also common in duplicated genomes. These are polymorphic variations that differ between allelic copies, segregates among generations, thus are a source informative markers for quantitative genetics and population studies. The biggest challenge in duplicated genomes is discriminating SNPs from those of other sequence variants (PSVs, MSVs and SNPs) as all seem to appear as polymorphic sites thus complicating any genetic analysis (Sánchez et al., 2011).

The Atlantic salmon genome along with other members of the Salmonidae family is in the process of reverting back into a stable diploid state, through deleterious mutations, gene silencing or by losing the redundant segments of the genome. However, extant salmonids have not fully completed the re-diploidisation process yet. Half of the

salmonids genome has been estimated to still be in a duplicate form (Allendorf, 1978). This is evident with widely varying numbers of haploid chromosomes in the Salmonidae family ranging from 26 to 51, while closely related teleosts have relatively stable haploid chromosome numbers around 24-25 (Naruse et al., 2004) eg: Northern pike (*Esox lucius*) possess 25 chromosomes, a member of the closest related diploid sister group to the Salmonidae family (Rondeau et al., 2014). The European Atlantic salmon populations typically have 29 chromosome pairs with 74 chromosome arms while North American populations generally have 27 chromosome pairs with 72 chromosome arms (Lubieniecki et al., 2010 and refs cited therein). The salmonid family shows that keeping chromosome numbers stable is not a prerequisite post-WGD. Rather most species possess a reduced number of chromosomes from the duplicated number right after WGD (see Fig. S1 in Glasauer & Neuhauss, 2014). Although WGD doubles up the chromosome set of an organism, chromosomes go through dynamic rearrangements, one of which is Robertsonian translocations. These centric fusions result in two chromosomes fusing at centromeric regions thus reducing the number of chromosome pairs. Such fusions also largely explain the common existence of meta- and acro-centric chromosomes observed among salmonids (Allendorf et al., 2015; Wright et al., 1983). In addition, sex-specific tetraploid segregation is a well-known phenomenon in the Atlantic salmon where all loci segregate in traditional Mendelian fashion (in diploid form) in females while males represent residual tetrasomic inheritance (Danzmann & Gharbi, 2001). This observation was in agreement with the multivalent pairing of chromosomes during meiosis (Timusk et al., 2011). Many linkage maps constructed so far, in various members of Salmonidae family, highlighted a marked difference between the sex and the average number of crossover. An almost equal distribution of crossovers is observed along the female chromosomes enabling

reliable estimates of recombination frequencies and the position of many markers within linkage groups. In contrast males display telomere-specific recombination patterns so their genetic linkage maps have poor resolution in centromeric regions thus generate shorter overall maps (Gharbi et al., 2006; Gonen et al., 2014; Lien et al., 2011). Straightforward analysis of the Atlantic salmon genome would thus be challenging, if not impossible, given the complexities associated with WGD and the many chromosomal rearrangements.

The international collaboration to sequence the Atlantic salmon genome (ICSASG) announced its plan to sequence the genome of the Atlantic salmon as a model and a representative of the Salmonidae family in 2010 (Davidson et al., 2010). A year later, the first assembly (*Ssal_v1*; ASM23337v1) was made publicly available in October 2011 on GenBank (NCBI). This version of the assembly involved only about a 6th of the expected genome size of the species (see Results section 5.4.5). The committee improved the assembly by applying a hybrid model of Sanger sequencing, Illumina short reads and PacBio long reads for scaffolding. This approach has significantly improved the second version of the assembly (*Ssal_v2*; GCA_000233375.4) comprising 965,912 contigs with N50 contig length of 36kb. In addition to that the assembly was presented as chromosomes as opposed to previous version of scaffolding. Finally the third and the most updated version of the Atlantic salmon assembly (*Ssal_v4*; AGKD000000000.4) was made publicly available in June 2015 on GenBank using both genome (WGS) and transcriptome (TSA) assembly sequences. Linkage mapping was used to position the scaffolds into 29 single chromosome sequences (Lien et al., 2016). The genome length of the current assembly (*Ssal_v4*) was 3.4×10^9 closest to the estimated genome size of the species. The total number of contigs was 839,389 and they applied a similar approach to the hybrid assembly. The genome of the Atlantic

salmon has long been known to possess big fragments of repetitive sequences, reported as 60% (Lien et al., 2016) one of the highest repeat content observed in any vertebrate (McCluskey & Postlethwait, 2014). Due to the nature of such fragments the majority could not be anchored to a specific chromosome thus still represents complications for the following downstream analysis (Lien et al., 2016).

One of the fundamental prerequisite of science has always been trying to achieving highly reproducible results which in the case of animal experimentation can lead to the use of large numbers of animals to give statistically relevant results because of the natural variability in many biological systems. Isogenic clonal lines offer the potential to reduce the number of animals required to produce significant results with their unique genotype of increased genetic uniformity (see Chapter 1). Such lines can be produced in two subsequent generations in fish either through gyno- or andro-genesis (see review by Komen & Thorgaard, 2007) as opposed to successive mating of siblings for over 20 generations as in mice or rodents, to produce so called inbred lines (Beck et al., 2000). Both techniques require applying chromosome set manipulations to inactivate the DNA content of one of the parents (by irradiation of the egg / sperm nuclear content respectively in the case of androgenesis and gynogenesis). Then diploidy is restored in the embryo by applying a heat / pressure shock to suppress endomitosis resulting in the retention of two identical copies of the male or female haploid genome, resulting in the first generation doubled haploids (DHs) (see Chp 1 for the detailed explanation of each technique). A sib-group of these fish will all be slightly different as they will have undergone different recombination events during meiosis but all will be homozygous at all loci. The overall survival and robustness of such fish can be highly variable and the ongrowing process needs to be optimised to minimise mortality. In order to produce the second generation of isogenic fish we can only use

fish that develop and reach sexual maturity. A secondary consequence of producing homozygous fish is that this may result in single sex offspring if the parental fish had a homogametic sex-determination system. In the case of Atlantic salmon the female is the homogametic sex (XX) and gynogenetic offspring will be all-female. In this case it will be possible to utilise meiotic gynogenesis to produce the second generation DH isogenic lines, although there will be recombination as all loci are homozygous this will not induce any heterozygosity in the offspring from a single parent.

Such isogenic lines are of interest in aquaculture related research to elucidate the genetic variation due to individual alleles at a locus. Since there is no heterozygosity in both G1 and G2 progenies, every allele is expressed in homozygous form. (e.g: A heterozygote “Aa” suppresses the expression of a recessive trait due to dominance effect, such cases are eliminated in DH individuals as they will be 100% homozygous thus denoted as “AA” or “aa”). This doubled additive genetic variance in DH individuals reveals extreme genotypes, particularly the recessive alleles (Bongers et al., 1997b). Festing & Altman, (2002) suggested the best strategy for working with either inbred or isogenic clonal lines would be to use a small numbers of animals from several strains or lines to ensure high statistical power rather than using a massive number of animals coming from different origins. This not only creates noise in the dataset but also requires a massive amount of fish to reach decent statistical power. As of today, most regulatory bodies at national levels (e.g: UK government Home Office) disseminate the advantages of **3R** (**R**educe, **R**eplace, **R**efine) framework with the Act 1986 (Animal Scientific Procedures). Grimholt et al. (2009) critically reviewed the need for genetically standardised lines in the Atlantic salmon for research purposes. Isogenic clonal lines have successfully been produced in rainbow trout (Quillet et al., 2007), Nile tilapia (Hussain et al., 1998; Muller-Belecke & Horstgen-Schwark, 2000), zebrafish

(accelerated with Streisinger et al., 1981 protocol eg: Mizgireuv & Revskoy, 2006 for cancer research and the most recent study on isogenic line production via androgenesis by Hou et al., 2015), medaka (Naruse et al., 1985), common carp (Bongers et al., 1997a), ayu (Han et al., 1991; Taniguchi et al., 1994), Japanese flounder (Hara et al., 1993), olive flounder (Yamamoto, 1999) and red seabream (Kato et al., 2002).

Advances achieved in next generation sequencing technologies have revolutionised the entire experimental design for all sorts of genetic studies. Such technologies are capable of generating large numbers of SNP markers that are well distributed along the genome of interest. One of the most popular approaches is called double digest restriction-site associated DNA sequencing (ddRADseq), first described by Peterson et al. (2012) and has been applied to many species even without a reference genome being available (review: Elmer & Meyer, 2011). ddRADseq is a genome complexity reduction technique where enzymatically fragmented genomic DNA is barcoded and pooled from many individuals (family or population wide) at the flanking regions of restriction enzyme recognition sites. This not only enables reliable SNP calling but also ensures true base calling with the deeper sequencing available. Although NGS technologies were initially designed to sequence diploid organism (Human Genome Project), such platforms can also deal with organisms of duplicated origin or polyploids as in plant genomics. The literature involves growing numbers of studies utilising high-throughput data generation power of NGS in duplicated genomes in conjunction with stringent filtering criteria to remove non-allelic forms of the sequence variants. For example, Gonen et al. (2014) used the RADseq platform and applied a stringent filtering criterion (het>70% removed) to remove most of the paralogous sequence variants from the dataset to be used for genetic linkage mapping in Atlantic salmon. A recent high-density SNP chip study involved implementing the same strategy to remove excess

heterozygotes from the genotyping array in the Atlantic salmon (Houston et al., 2014). Similarly, Hohenlohe et al., (2011) applied high filtering procedures to a dataset generated by RADseq to deal with duplicated genomes of salmonid species in both rainbow and cutthroat trout.

In this study, we aimed to verify an optimised isogenic clonal line production protocol in the Atlantic salmon, a species of prime commercial interest which has already been a target of several studies on the induction of meiotic and mitotic gynogenesis (Johnstone and Stet, 1995) and polyploidy (Johnstone, 1992; Smedley et al., 2016). The ddRADseq technique was used as an improved way of detecting any residual contribution from the irradiated sperm. Considering the duplicated genome of the Atlantic salmon, we utilised well-established genomic resources for the species (reference genome assembly) to remove all duplicated loci, redundant copies of the genes post-WGD and repetitive elements following *de novo* assembly. Unlike the previous studies, we applied BLAST (NCBI) approach, as opposed to assuming excessive heterozygotes as being fixed sites in previous studies, thus eliminated the possibility of using multi-copy loci for the investigation of sire contribution in both G1 and G2 families. This approach allowed us to produce reliable results by identifying and subsequently discriminating one-copy loci from those of duplicated ones. Overall, this study represents a well-defined pilot study that provides evidence on how high-throughput sequencing technologies can be applied to the analysis of duplicated genomes, providing there is a good quality reference genome assembly available for the species of interest. This will ensure the subsequent establishment of reliable techniques to be used by future generations to propagate more lines of interest.

5.2 Materials and methods

5.2.1 Production of Isogenic clonal fish lines

5.2.1.1 Overview

The production of putative clone founders (G1) and putative clonal lines (G2) used in the present study were carried out in Norway, IMR (Bergen). In the first generation, to propagate homozygous clone founders, a single male and a single female were used as outbred parental fish while in the subsequent generations, to propagate putative clones, sperm pool consisting of 2 males and each survivor female from G1 progeny were used (see Figure 5.1 describing experimental design). Mitotic gynogenesis was applied in the first generation to produce putative homozygous clone founder while meiotic gynogenesis was applied in the subsequent generation, due to the higher survival level normally achieved, to produce putative clonal fish lines each coming from a different G1 fish. The initial study, however, was carried out with a higher numbers of dams to establish isogenic clonal fish lines, first time ever in the Atlantic salmon under an EU project, AquaExcel (FP7) extended to AquaExcel²⁰²⁰ (see deliverables of the project online). In total >2100 G1 fish were produced by IMR (240 from 2011/12 year class; 800 from 2012/13 year class; 1100 from 2013/14 year class) in a facility that had been specifically established for these fish. Each family was produced by artificial fertilisation of the eggs within two groups: i. UV irradiated sperm (for mitotic gynogenetics) and ii. normal milt (as bi-parental control groups). After eight months the surviving fish were PIT tagged and a small fin clip was taken for genotyping. Initial genotyping was carried out in Norway using 18 microsatellite markers (Glover et al., 2009).

In total, 33 samples (Figure 5.1) were provided to University of Stirling for the verification study. The samples included; outbred founders (provided as genomic DNA), six doubled-haploids (G1 fish, provided as fin tissues) and five second generation doubled-haploids from each of the above G1 fish (one of the G1 fish did not have progeny) so in total, twenty-five fish have been received from the putative isogenic clonal fish lines. Once putative isogenic clonal lines reach up to 500dd-800dd (Larva with yolk sac provided; after hatching before the first feeding stage) the head part of the juveniles were sent to Stirling for further analysis.

The library pool also included a haploid family (300dday) with ten progeny from the previous experiments carried out in Stirling, UK on optimisation of the UV irradiation in salmon as well as parents provided by Landcatch, UK. In total, 45 fish were used for ddRAD sequencing. Outbred parents were triplicated while all members of G1 and G2 families were duplicated in the ddRAD library. Pedigree information is given in Table 5.1 (S1 Table includes all the details regarding to samples used in the present study).

5.2.1.2 *UV irradiation of sperm and pressure shock*

Sperm dilution was undertaken as follows: 4 ml of milt from one male salmon was mixed with 160 ml milt fluid (milt from several males were centrifuged until clear milt fluid was achieved then it was frozen and stored at -20 °C; thawed before the experiment). In total twelve 15 ml aliquots of the diluted milt were irradiated under the UV light for 6 or 8 mins in 8 cm Petri dishes at 480 $\mu\text{W}\cdot\text{cm}^{-2}$. Following irradiation of diluted sperm, this was transferred to 25 ml polyethylene (PE) containers and stored refrigerated and in darkness until fertilization. In total, 1000 salmon eggs from one female were artificially fertilized with each of the sperm aliquots and left to hydrate in 0.5 L PE bottles at 8 °C until pressure treatment. Bi-parental control groups received the

same procedure without the UV irradiation of sperm and without pressure shock (as ordinary fertilisation).

A pressure shock of 9500 psi for duration of 5 minutes was applied to each group at 4600 and 4800 min°C transferred into pressure chambers from the PE bottles. Survival checks were carried out at the eyed stage and at hatching. After eight months, surviving fish were PIT-tagged and fin clipped to be used for genotyping. After sampling all fish were returned to the tanks for further on-growing under 7/24 controlled environment. Initial genotyping was carried out using 18 microsatellite markers (Glover et al., 2009). The physical positions of microsatellites were identified on the genome assembly.

5.2.2 ddRAD library preparation and sequencing

DNA extractions were carried using a universal salt buffer includes SSTNE-SDS as detailed explained by (Grant et al., 2016). DNA concentration and purity of each sample were assessed by using a Nanodrop spectrophotometer and the molecular weight of DNA was assessed by agarose gel electrophoresis. Each sample was diluted to a concentration of 10 ng/ μ L in 5 mM Tris, pH 8.5. Final DNA concentration assessment of all samples was completed in a customized well plate in a Quantica qPCR thermal cycler (Techne, UK) using a dsDNA fluorescent dye, Qubit® dsDNA HS Assay Kit (Invitrogen, UK) prior to the ddRAD library construction.

The ddRAD library preparation protocol followed essentially the methodology originally described in Peterson et al. (2012) with slight modifications explained by (Palaiokostas et al., 2015b). Early pooling of the samples following the restriction digestion and ligation of the adapters and enrichment of the entire library was the main modification in the protocol used here as opposed to individual enrichment of each sample and latter pooling in the original protocol.

Each sample (0.021 µg DNA) was digested at 37°C for 90 minutes with SbfI (rare cutter, recognising the CCTGCA|GG motif) and SphI (common cutter, recognising GCATG|C motif) high fidelity restriction enzymes (New England Biolabs; NEB), using 30U each enzyme per microgram of genomic DNA in 1× CutSmart™ Buffer (NEB). No heat inactivation was performed. The restriction digestion reaction volume per individual sample was 6 µL. Individual-specific combinations of P1 and P2 adapters (Table S1, available online), each with either a unique 5 bp or 7 bp barcode (in order to avoid “registration” issues on the Illumina sequencer), were ligated to the RE fragmented DNA at 22 °C for over three hours by adding 0.6 µL 100 nmol/L adapters, 0.15 µL 100 mmol/L rATP (Promega), 0.25 µL 10× CutSmart™ Buffer (NEB), 0.12 µL T4 ligase (NEB, 2 M U/mL) and reaction volumes made up to 12 µL with nuclease-free water for each sample. Regarding registration issues, Illumina uses a green laser to read G/T bases and a red laser to read A/C. At each cycle of sequencing – especially the first 10-15 bases read when clusters are identified - at least one of the two nucleotides for each colour channel needs to be read to ensure proper registration. The barcodes were selected 1) to differ from each other by at least 3 bases, and 2) to balance green/red laser detection at each base position. However, each ddRAD fragment has an area of low complexity – the RE site itself. In order to compensate for this – half the barcodes were 5-bases, the other half were 7-bases long. This frameshifts the reads such that balanced green and red laser detections occur across this low complexity region. Following ligation reactions all samples were combined in a single pool (for one sequencing lane) and purified by MinElute PCR clean up kit (Qiagen, Invitrogen).

Size selection (416-706 bp) was performed by agarose gel separation and was followed by gel purification and PCR amplification. A total of 50 µl of the amplified library (13 cycles) was purified using an equal volume of AMPure beads. After eluting into 20 µL

EB buffer (MinElute Gel Purification Kit, Qiagen), the library was quantified before and after bulk PCR amplification by Qubit (dsDNA HS, Invitrogen). The ddRAD library was then diluted down to 9.5 nM final library concentration by using freshly prepared 0.2M NaOH / 1% Tween 20 (Alpha Laboratories). Denaturation of final library was achieved by using both chemical (NaOH) and heat treatment (2 minutes incubation at 95°C) according to Illumina's protocol. PhiX (control) library (Illumina) and the ddRAD library were mixed to reach the desired concentration of 10.25 pM and the final library was loaded onto a Miseq (Illumina) V2-300 kit for paired-end sequencing.

5.2.3 Microsatellites

Genotyping with microsatellites was carried out with a panel of 11 fluorescently-labelled loci (Cp 2.2; Appendix) as a means of double checking ddRAD sequencing results where varying levels of sire contribution was observed in five of the putative clonal families (G2). The microsatellite loci (Vasemägi et al., 2005) were selected initially for their wide usage and reliability for an another project which aims to screen Atlantic salmon broodstock for genetic variation.

Reactions consisted of a total volume of 10µl, comprising half volume coming from MyTag™ 2x Mastermix (Bioline, UK) solution ensures 1X concentration reaction at the end, 3.2µl distilled water, 1µl DNA (25ng/µl), and 0.2µl of 1uM tailed forward primer (see Chapter 2, 2.7.2 for the details of fluorescent primer tailing approach), 0.3µl of 10uM of non-tailed reverse primer and 0.3 µl of 10uM fluorescent dye. PCR reactions were conducted on a Biometra TGradient thermal cycler that was programmed with a 1 minute denaturation step at 95°C, 95°C denaturation for 15 seconds; 60°C annealing for

20s and 72°C extension for 30 s for 32 cycles, without requiring a final extension for all PCR multiplexes 1 - 3.

Beckman-Coulter CEQ8000 sequencer and associated software was used for size determination of the fluorescently labelled PCR products. Each row of 8 samples ran for 45 min using Beckman Frag-3 (size range 60-400bp) genotyping method for multiplex_2 and multiplex_3 while multiplex-1 was genotyped by using Frag-4 (size range 60-600bp) genotyping method. Allele scores were manually verified.

Table 5.1: The pedigree of the samples

Sample ID	Family source	Received labels	Clonal line labels
MO-863	Outbred sire	Outbred dad	n/a
MO-864	Outbred dam	Outbred mum	n/a
MO-855	G1	DH91	2.1 to 2.5
MO-856	G1	DH93	3.1 to 3.5
MO-857	G1	DH115	No progeny
MO-858	G1	DH133	4.1 to 4.5
MO-859	G1	DH154	5.1 to 5.5
MO-860	G1	DH224	1.1 to 1.5
MO-867	G2	2.1	DH1 fam
MO-868	G2	2.2	DH1 fam
MO-869	G2	2.3	DH1 fam
MO-870	G2	2.4	DH1 fam
MO-871	G2	2.5	DH1 fam
MO-872	G2	3.1	DH2 fam
MO-873	G2	3.2	DH2 fam
MO-874	G2	3.3	DH2 fam
MO-875	G2	3.4	DH2 fam
MO-876	G2	3.5	DH2 fam
MO-877	G2	4.1	DH3 fam
MO-878	G2	4.2	DH3 fam
MO-879	G2	4.3	DH3 fam
MO-880	G2	4.4	DH3 fam
MO-881	G2	4.5	DH3 fam
MO-882	G2	5.1	DH4 fam
MO-883	G2	5.2	DH4 fam
MO-884	G2	5.3	DH4 fam
MO-885	G2	5.4	DH4 fam
MO-886	G2	5.5	DH4 fam
MO-887	G2	1.1	DH5 fam
MO-888	G2	1.2	DH5 fam
MO-889	G2	1.3	DH5 fam
MO-890	G2	1.4	DH5 fam
MO-891	G2	1.5	DH5 fam
MO-865	Haploid sire	B04	Haploid parents
MO-866	Haploid dam	C	Haploid parents
MO-892	Haploid	Offspring-1	Haploid progeny
MO-893	Haploid	Offspring-2	Haploid progeny
MO-894	Haploid	Offspring-3	Haploid progeny
MO-895	Haploid	Offspring-4	Haploid progeny
MO-896	Haploid	Offspring-5	Haploid progeny
MO-897	Haploid	Offspring-6	Haploid progeny
MO-898	Haploid	Offspring-7	Haploid progeny
MO-899	Haploid	Offspring-8	Haploid progeny
MO-900	Haploid	Offspring-9	Haploid progeny
MO-901	Haploid	Offspring-10	Haploid progeny

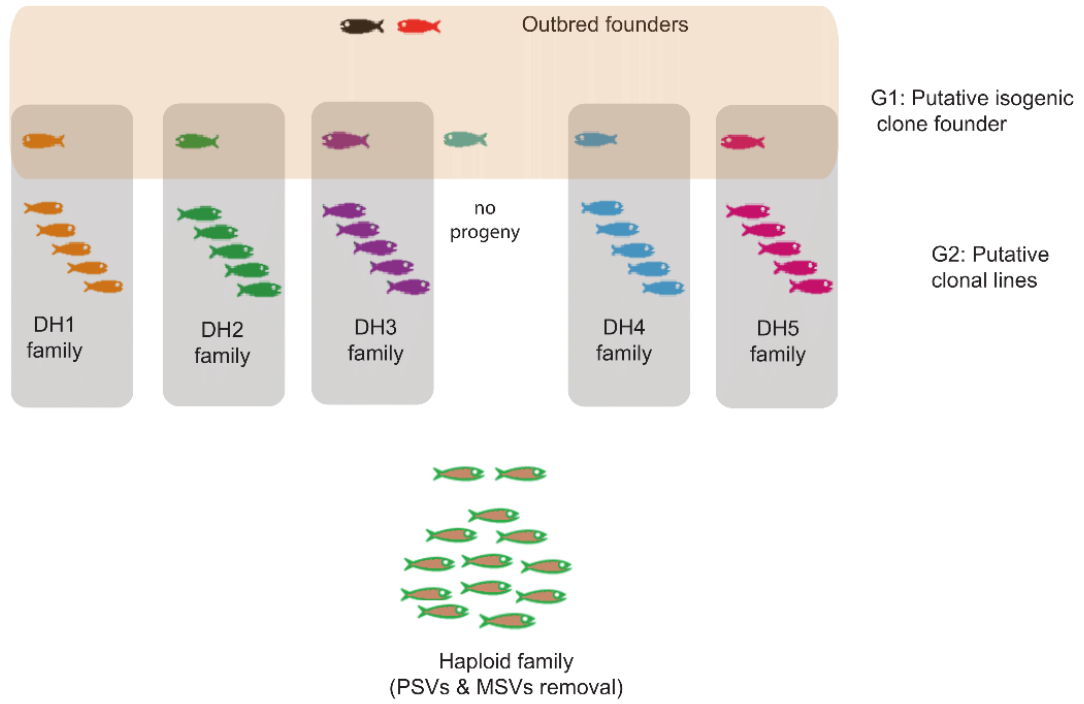


Figure 5.1: The schematic diagram of experimental design used in the present study.

5.3 Data Analysis

5.3.1 Sequence Quality Control (QC)

Initial quality control checks of the raw data were carried out by FastQC v. 0.11.3 (Andrews, 2010) for assessment of sequence quality scores (Phred30), GC/AT content and over-represented sequences. Reads of low quality (Phred score under 30), missing the restriction site or with ambiguous barcodes were discarded by using *process_radtags* module implemented in Stacks (Catchen et al, 2011). This module also demultiplexes data. The filtered read files were then renamed to reflect sample names, barcodes were trimmed. Retained reads were trimmed to a length of 135bp due to quality score of paired-end reads (P2) dropped to the calls of *reasonable quality* indicated with orange background (Phred score 20-28) instead of *very good quality calls* indicated as green (Phred score >30) (see Report S1 & S2 respectively for both P1 and P2 raw reads QC report: compare the outcome of per base sequence quality graphs, available online).

5.3.2 Genotyping ddRAD alleles

The trimmed reads were sorted into loci and genotyped using the Stacks pipeline v1.30 (Catchen et al, 2011). The likelihood-based SNP-calling algorithm (Hohenlohe et al., 2011) implemented in Stacks evaluates each nucleotide position in every ddRAD-tag of all individuals, thereby differentiating true SNPs from sequencing errors. The core modules ('ustacks', 'ctsacks' and 'sstacks') of the Stacks pipeline were used to process all reads. Processing of 45 individuals was performed by generating *de novo assembly* of the sampled loci. Each family was analysed in separate batches in the same catalogue. The sequence for each ddRAD allele began at a defined position of the enzymes recognition site: "TGCA" motif for the *SbfI* cut site and "CATG" motif for

SphI cut site in the catalogue after removing barcodes. In order to maximise the number of informative markers and minimise the amount of missing or erroneous data, only ddRAD-tags retrieved in at least 70% of the samples in each family, carrying up to three SNPs as well as four alleles were used (Figure 5.2).

5.3.3 SNP calling in G1 and G2 families

A minimum stack depth of at least 10 and a maximum of 2 mismatches were allowed in a locus in an individual, and an additional mismatch was allowed between individuals for G1 and G2 families. Highly repetitive ddRAD tags were removed from the final data set by applying -t option in the *ustaks* program. Parameters for each module were left at their default values except for the number of mismatches allowed between loci when building the catalogue (-n = 1). By increasing the -m and -n parameters from their default settings to (-m = 10 and -n = 1), SNP calling confidence was increased while missing data was minimised.

5.3.4 SNP calling in haploid family

Haploid analysis involved specifying a distance of zero (-M = 0) between stacks when processing a single individual along with disabling calling of haplotypes from secondary reads (-H). This was taken into special account in the case of having enough secondary reads propagated from PCR or sequencing errors as false positive stacks might give rise to a second allele at a locus in an individual. Such cases were avoided by applying -m, -M, -n parameters set to 10_0_1 as well as -t and -H (disabling genotype call from secondary reads) in haploid family. This family required resetting stacks parameters to -m, -M, -n to 4_0_1 as well as -t and -H applied for the final analysis (see Discussion).

5.3.5 Distribution of polymorphic ddRAD loci

Within the full set of ddRAD loci the following were excluded: (i) any alleles missing in both parents, (ii) alleles that were represented in less than 70% of the progeny. Table 5.2 shows a filtered set of four allelic markers carrying up to three SNPs showing presence / absence segregation pattern in the offspring indicating dam alleles. Initial examination of any variants (heterozygotes) was carried out on the web interface of the Stacks pipeline. After the filters were applied all parental genotypes were carefully inspected, particular attention was paid to the ones where dam and sire had distinctive alleles. After initial examination, the main analysis was carried out on the filtered genotype files, in Excel. In the case of observing any heterozygotes in the progeny both in putative clone founder (G1) and putative clonal fish (G2) each locus was cross-checked on the web interface for the technical details of a valid locus. These were defined as having an allele depth (>10 per each allele), having clear stacks to make a genotype call and not-involving a repetitive motif which might be misleading and give erroneous results.

5.3.6 Initial investigation of putative sire contributor loci

The existence of potential residual chromosomal fragments in resulting offspring in both G1 and G2 families was checked by examining the segregation patterns of alleles, briefly, by looking at presence/absence of alleles in each of the genotypes. First, all markers were checked where the dam was homozygous and the sire was heterozygous or homozygous for a different alleles compared to the dam to detect any sire contribution. Those showing any allele from the sire were categorised as potential paternal contributor loci, PPCL (see Table 5.2). Then a random selection of PPCL was further investigated using NCBI-Blastn tool. In total, 50 (out of 127 PPCL) randomly

picked loci from each parental genotype (see Table 5.2) in G1 family were blasted to the whole genome shotgun sequencing of salmon genome while 20 randomly picked loci were further investigated in each of the G2 families (100 loci in total).

5.3.7 Identification of multi-copy loci

Having observed high levels of heterozygosity (see Table 5.2) caused by either sire contribution or homologous loci due to WGD in salmon, the best approach was to remove all multi-copy loci for the verification of isogeny in G1 and G2 families respectively. For that, all polymorphic ddRAD alleles before filtering (8548 loci in total) were pooled after individually labelled (marker ID combined with family info eg: 1259_G1) and Blasted against all three available versions (*Ssal_v1*; *Ssal_v2* and *Ssal_v4*) of the Atlantic salmon genome assembly. Perl scripts are available in Supplementary Material 1.

The results of the Blast search was filtered based on i) e^{-40} and smaller, ii) alignment length 130bp and higher (which ensures 96.2% similarity rate considering initial fragments were 135bp after trimming) and iii) more than 10 mismatches were removed from the dataset.

5.3.8 Investigation of sire contribution with one-copy loci

Following identification and subsequently removal of the multi-copy loci, each family was further investigated for potential sire contribution to progeny by using only non-duplicated loci. First, each of these loci were checked for the segregation of the alleles on the stacks web interface recording genotypes of parents and the progeny observed on an Excel sheet. This data was translated into a tabular format, including all the genotypes (see Table 5.7 for demonstration purposes). However some parental

genotypes with common alleles were difficult to interpret whether the shared allele was segregating from sire or dam (e.g: parental aa/ab (♂ / ♀) genotype is expected to give rise “aa, bb” genotypes after application of mitotic gynogenesis. However due to “a” allele being shared both by sire and the dam it is not clear where the “a” allele is coming from. Such cases were categorised as only half informative markers). There were also cases where the female genotype was missing (Table 5.7). Such cases were eliminated as non-informative nature of the genotypes. Likewise in the case of both sire and dam sharing the same genotype, these were removed from the summary table after inspection for any heterozygosity observed (eg: parental ab/ab (♂/♀) genotype is expected to give rise “aa, bb” genotypes. Any heterozygosity observed in the progeny would indicate a failed mitotic gynogenesis protocol). After all, an easier format (including only informative and half informative markers) was selected to visualise any sire contribution to progeny for each family (Tables 5.8; 5.9 and 5.10).

5.3.9 Finding the position of PCR primers using NCBI-BLAST

Given the microsatellites used were not initially designed specifically for the purpose of this study, it was essential to locate the position of each locus on the genome to ensure the microsatellite loci were informative. However, due to the heuristic nature of Blast, queries involving short sequences such as primers often return incomplete data. This was significantly improved by concatenating two primer sequences: forward and reverse by adding an extra 20 N nucleotides in between (Integrated DNA Technologies). The orientations of stings were not initially taken into consideration given Blast searches use both strands for the matches. Species name was specified in Blastn module. A series of adjustments on both program selection and algorithm parameters were applied. Programme selection was optimised for *somewhat similar sequences*. The *low complexity filter* was turned off while *expect threshold* was

increased to 1000 and word size decreased down to 15. This returned successfully the position of the microsatellites on genome as well as expected fragment size (S2 Table).

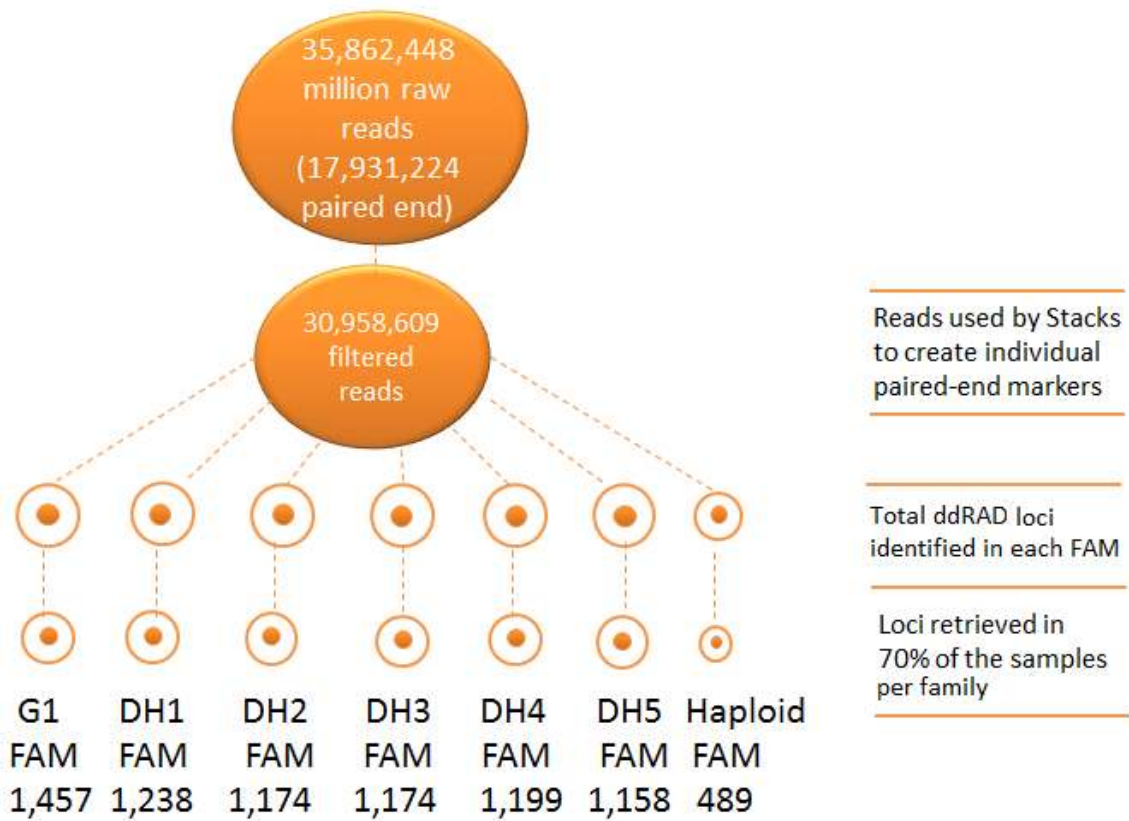


Figure 5.2: Sequencing and ddRAD-tag summary. Details of the number of reads before and after filters (orange disk) followed by the reconstructed number of polymorphic ddRAD loci after filtering.

5.4 Results

5.4.1 ddRAD sequencing

A total of 35,862,448 (each 162 bp long) reads were obtained (17,931,224 paired-end) from the sequencing of the 46 individuals including outbred parents, G1 family, five each of the G2 families and haploid family (Figure 5.2). Following Stacks pipeline filtering low quality reads (Phred33, quality score under 30) (229,089), ambiguous barcodes (4,419,238), ambiguous RADtags (255,512) were removed which subsequently resulted in 86.3% of the raw reads being retained, corresponding to 30,958,609 reads. Raw reads of the data from this study were deposited in the European Bioinformatics Institute (EBI)-SRA database with the unique accession number of ERP011576.

5.4.2 Distribution of the ddRAD alleles

In total, 1457 polymorphic ddRAD loci were detected in G1 family while an average of 1188 polymorphic ddRAD loci were observed in G2 putative clone families (Table 5.2) and 489 polymorphic ddRAD loci were retained in the haploid family (Table 5.3). This variation observed among the families was simply due to different levels of replication of samples in ddRAD library preparation (Figure 5.2, S1 Table). Outbred parents (parents of G1 family) were triplicated randomly in the library to ensure high read depth for parental genotypes to be matched with offspring later during the bioinformatics analysis. Offspring of G1 family (6 fish of which 5 were the parents of G2 families, one fish did not have progeny) and G2 (25 fish, 5 per family) were duplicated. The DNA samples of the haploid family was duplicated in parents, while offspring were used only once in ddRAD library construction thus produced almost three times fewer reads compared to G1 and G2 families (see Figure 5.2).

The polymorphic ddRAD loci were analysed for any sire contribution to the progeny in each family. In total, 1457 ddRAD loci were observed in the G1 family of which 127 of the loci (8.7% of the total loci) initially appeared to show paternal contribution to the offspring based on a presence/absence segregation pattern (see Table 5.2). The frequency of potential paternal contributor loci increased to 26.3%, 23%, 27.3%, 28% and 22.6% respectively in the subsequent generation of clonal line production in the five G2 families (Table 5.2). In total, 489 polymorphic ddRAD loci were detected in the haploid family in which 143 of them (29%) appeared to possess residual sire contribution in the progeny (see Table 5.3).

5.4.3 Initial investigation of the putative sire contribution in duplicated genome of salmon

The frequency of homolog loci among the potential paternal contributor loci was similar in both G1 and G2 families. Out of 50 randomly picked loci from each parental genotype in the G1 family, the frequency of homolog loci was 60% and 40% was repetitive elements. Homolog loci were easy to identify with completely identical fragments in the genome assembly coming from two different contigs while repetitive elements would return many contigs with high similarity rate in the whole genome shotgun of the Atlantic salmon in gene bank (NCBI). Within the five G2 families a further 20 loci were studied in each family (100 loci in total), 44% were homologous loci 37% were repetitive elements and 19% were single copy genes in the genome assembly of salmon. So as to double check the existence of repetitive elements in both G1 and G2 families the loci showed the frequency of 40% and 37% respectively was further blasted into *ref_seq* (well annotated genomic database-NCBI) and various sequence similarities were observed mainly in pike genome (*Esox lucius*) – the most closely related species to the salmonids – as well as all sorts of other teleost. Given the

cases further investigated initially, there was no convincing sign of paternal contribution in G1 family but in 19% of the times only one contig carrying given allele which does not match up with maternal allele was likely indicating a potential sire contribution among G2 families.

Table 5.2: Distribution of ddRAD alleles in each family. The first genotype refers to sire and the second is to dam.

Parental Genotype	G1 Family			G2 Families														
	U.tags ¹	PPCL	% PPCL	DH1 Family			DH2 Family			DH3 Family			DH4 Family			DH5 Family		
				U.tags ¹	PPCL	% PPCL	U.tags ¹	PPCL	% PPCL	U.tags ¹	PPCL	% PPCL	U.tags ¹	PPCL	% PPCL	U.tags ¹	PPCL	% PPCL
aa/bb	175	1	0.6	345	93	27.0	336	87	25.9	352	88	25.0	367	88	24.0	360	51	14.2
ab/UNK	18	0	0.0	116	0	0.0	127	0	0.0	49	0	0.0	15	0	0.0	0	0	0.0
UNK/ab	21	0	0.0	0	0	0.0	1	0	0.0	0	0	0.0	15	0	0.0	10	0	0.0
aa/ab	445	11	2.5	30	19	63.3	20	14	70.0	28	20	71.4	29	24	82.8	22	21	95.5
ab/aa	431	13	3.0	627	121	19.3	576	80	13.9	618	106	17.2	637	110	17.3	636	80	12.6
ab/ac	40	8	20.0	7	6	85.7	7	5	71.4	10	9	90.0	13	11	84.6	9	8	88.9
ab/cd	1	0	0.0	0	0	0.0	1	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0
ab/cc	7	1	14.3	24	2	8.3	21	4	19.0	23	7	30.4	25	7	28.0	21	2	9.5
cc/ab	5	0	0.0	0	0	0.0	2	1	50.0	1	1	100.0	1	0	0.0	0	0	0.0
ab/ab	314	93	29.6	89	84	94.4	83	79	95.2	93	89	95.7	97	96	99.0	100	100	100.0
TOTAL	1457	127	8.7	1238	325	26.3	1174	270	23.0	1174	320	27.3	1199	336	28.0	1158	262	22.6

¹: Unique tags represent the total number of loci in given genotype

²: PPCL, Potential Paternal Contributor Loci

Table 5.3: Distribution of ddRAD alleles in haploid family. The first genotype refers to sire and the second is to dam.

Haploid family			
Parental Genotype	U.tags ¹	PPCL ²	% PPCL
aa/bb	21	15	71
ab/UNK	194	0	0
UNK/ab	4	0	0
aa/ab	49	7	14
ab/aa	141	79	56
ab/ac	4	3	75
ab/cd	0	0	0
ab/cc	2	2	100
cc/ab	0	0	0
ab/ab	74	37	50
TOTAL	489	143	29%

5.4.4 Removal of multi-copy loci

Blast analyses carried out with all polymorphic ddRAD alleles pooled from all families against the different genome assemblies (NCBI) gave a range of results. The *Ssal_v1* identified only 66% of the total polymorphic ddRAD loci within the assembly while *Ssal_v2* and *Ssal_v4* identified 99.43% and 99.46% respectively. A summary table is provided (Table 5.4). The most recent and complete assembly *Ssal_v4*, had genome length of 3.411.171.783 bp, closest to 3.2 Gb salmon genome size (Lien et al 2016) and was chosen as the most appropriate for the further investigation of sire contribution. (ca. 2.966.890.203 bp in *Ssal_v2* and 507.799.561 bp in *Ssal_v1*) for the identification and removal of multi copy loci (1815 non-duplicated loci were individually inspected).

The frequency of non-duplicated loci was similar among the three versions of the salmon genome assemblies, ranging between 18-22% (see Table 5.4). This was also observed at the family level, the frequency of non-duplicated loci ranged between 19-22%. For the detailed frequency distribution of the non-duplicated loci within each family and each assembly version, see Tables 5.5 and 5.6.

Table 5.4: Summary table of BLAST analysis among three versions of genome assemblies.

Genome assembly	Total hits found	Multi-copy loci*	Non-duplicated loci
<i>Ssal_v1</i>	5664 (66%)	4623	1041 (18%)
<i>Ssal_v2</i>	8500 (99.43%)	6580	1880 (22%)
<i>Ssal_v4</i>	8502 (99.46%)	6643	1815 (21%)

*: Any loci represented more than once in the genome assembly (including homologs and repetitive elements)

5.4.5 Comparison of three available genome assemblies (*Ssal_v1*, *Ssal_v2* and *Ssal_v4*) in salmon

A comparison of the outputs from the three available assemblies shows that *v1* was incomplete and only contained about a sixth of the genome size of the later versions. It is therefore not surprising that it identified only 66% of the polymorphic loci. The later versions (*Ssal_v2* and *Ssal_v4*) of the assembly significantly improved the number of hits found within each versions of the reference genome (Table 5.3) up to 99.43% and 99.46% respectively. Out of 8,548 total polymorphic ddRAD loci (pooled from G1 and G2 families) only 5,664 of them were found in the *Ssal_v1* while 8,500 and 8,502 hits were found in *Ssal_v2* and *Ssal_v4* assemblies. The limited number of loci (48 and 46 respectively) that could not be matched to any part of salmon genome within the newer versions was due to the nature of such fragments (Table 5.5). These fragments were tandem repeat microsatellites, thus could not be assigned to any parts of the genome. However, the 2884 fragments that could not been assigned in the earlier version of the genome assembly (*Ssal_v1*) was solely due to uncompleted nature of the assembly.

Table 5.5: Detailed BLAST analysis output of *Ssal_v1*, *Ssal_v2* and *Ssal_v4* genome assemblies (NCBI) in G1 and subsequent G2 families.

	<i>Ssal_v1</i> assembly						<i>Ssal_v2</i> assembly						<i>Ssal_v4</i> assembly					
	G1 family	G2 families					G1 family	G2 families					G1 family	G2 families				
	DH1 family	DH2 family	DH3 family	DH4 family	DH5 family		DH1 family	DH2 family	DH3 family	DH4 family	DH5 family		DH1 family	DH2 family	DH3 family	DH4 family	DH5 family	
Total ddRAD loci	1545	1295	1226	1283	1192	1165	1545	1295	1226	1283	1192	1165	1545	1295	1226	1283	1192	1165
Total hits found (%)	1028 (66%)	862 (66%)	805 (65%)	856 (66%)	788 (66%)	759 (65%)	1535 (99%)	1287 (99%)	1221 (99%)	1276 (99%)	1188 (99%)	1158 (99%)	1536 (99%)	1286 (99%)	1220 (99%)	1275 (99%)	1187 (99%)	1157 (99%)
No hits found ³	517	433	421	427	404	406	10	8	5	7	4	7	9	9	6	8	5	8
Duplicated loci ¹	837	708	661	697	643	617	1180	992	945	987	918	887	1194	1004	955	998	927	897
Non-duplicated loci ²	191 (13%)	154 (12%)	144 (12%)	159 (12%)	145 (12%)	142 (12%)	355 (23%)	295 (23%)	276 (23%)	289 (23%)	270 (23%)	271 (23%)	342 (22%)	282 (22%)	265 (22%)	277 (22%)	260 (22%)	260 (22%)
Filtered best hits⁴													334	277	262	273	256	257

¹: Some of these returned up to thousands times hits referring to existence of repetitive elements in the genome of salmon.

²: Total hits found in the genome assembly were subtracted by duplicated loci.

³: These are microsatellites, tandem repeats that could not be assigned in genome assembly.

⁴: These are non-duplicated best hits found, later screened for informative level of sire contribution.

Table 5.6: Detailed BLAST analysis output of *Ssal_v1*, *Ssal_v2* and *Ssal_v4* genome assemblies (NCBI) in haploid family.

	<i>Ssal_v1</i> assembly	<i>Ssal_v2</i> assembly	<i>Ssal_v4</i> assembly
Total ddRAD loci	843	843	843
Total hits found (%)	560 (66%)	835 (99%)	836 (99%)
No hits found ³	283	8	7
Duplicated loci ¹	460	671	676
Non-duplicated loci ²	100 (18%)	164 (19%)	160 (19%)
Filtered best hits⁴	97	158	156

5.4.6 Polymorphic one-copy ddRAD loci

In total, 333 one-copy loci were identified in G1 family (Table 5.7). Of those 10 loci had heterozygote genotype in the G1 family progeny (see Table 5.7 red coloured alleles). Detailed investigation of those revealed that these were due to secondary reads (see Discussion). Since such reads are more likely to carry sequencing errors, these were ignored in the G1 family. One locus (6659_G1) showing the “ab” genotype in the offspring was a suspicious case due to a long mono-nucleotide repeat prior to the SNP location, therefore ignored (1 locus out of 231 loci was heterozygous). Overall, the investigation of the non-duplicated loci provided no convincing evidence of any sire contribution in the G1 family at 230 loci which were either female informative due to distinctive allele set between parents or female heterogametic (see the difference between Table 5.7 and Table 5.8).

However in the G2 families varying levels of unknown alleles (not maternal) were detected in all G2 progeny (see summary Table 5.9). Since there were two males used and the milt was pooled to propagate all G2 families and no DNA was provided, the male genotype was denoted as “??” in G2 families (Table 5.9). Therefore any allele observed in progeny that did not match with female genotype was referred to as a potential sire allele. Removal of multi-copy loci and investigation of segregation pattern of only one-copy loci did not rule out the existence of non-maternal alleles in the progeny of putative clonal families. They showed a wide range of sire contributions between 10-25% among G2 families; 23%, 16%, 25%, 22% and 10% respectively (see Table 5.9 red coloured alleles). Furthermore the segregation of such non-maternal alleles was consistent among all progeny in each G2. This indicated that the sire contribution in the putative clonal progeny of G2 families is most likely due to a sub-

optimal UV irradiation treatment of milt. The presence of sire alleles in these G2 families rules out these families as possible isogenic clonal lines.

Regarding to results of the investigation of one-copy loci in the haploid family, Table 5.10 represents only informative markers following the removal of multi-copy loci in the haploid family. The sire contribution among the various types of informative markers was obvious. This was due to a sub-optimal protocol being used for haploid production in the Atlantic salmon. Thus haploids were not of use as was initially thought to identify and remove one source of sequence variants (PSV) observed in duplicated genomes.

Dataset S1 contains all one-copy polymorphic ddRAD loci that were used for the reliable verification of homozygosity of putative G1 and G2 fish in the Atlantic salmon. This dataset also involves the sequence of one-copy loci used for the verification of haploids.

Table 5.7: Detailed representation of all non-duplicated loci in G1 family (informative and non-informative markers, half informative markers where parents had a common allele and all other forms are listed). This table represented for demonstration purposes, subsequent tables, however, include a summary of only informative markers in both G2 families and haploid family.

G1 Family	♂ / ♀	Total number of loci*	Progeny carries only maternal allele	Progeny carries common alleles	PSC [#]
1. Informative markers	aa/bb	33	33	0	0
2. Female heterogametic markers I	aa/ab	100	9	90	1
3. Female informative markers	ab/cc	3	3	0	0
4. Female heterogametic markers II	cc/ab	1	1	0	0
5. Half informative markers	ab/aa	95	0	95	0
6. Less informative markers	ab/ab	67	0	58	8
7. Any other informative markers	--/ab	6	6	0	0
8. Any other non-informative markers	ab/--	2	0	2	0
9. Any other possible informative markers	ab/ac	8	4	3	1
10. No genotypes available, empty cells	--/--	18	0	0	0
Total		333	56	248	10

#: Potential Sire Contribution. The first genotype refers to sire and the second is from dam.

Table 5.8: The summary table of informative markers in G1 family (non-duplicated).

G1 Fam	♂	♀	Progeny	Total loci	♀ allele	♂ allele
1. 100% informative markers	aa	bb	bb	33	33	0
2. 100% informative markers	ab	cc	cc	3	3	0
3. Less informative markers	aa	ab	aa, bb	100	99	1
4. Less informative markers	ab	aa	aa	95	95	0
Total				231	230	1

Table 5.9: The summary table of informative markers in G2 families (non-duplicated)

Genotype	DH1 family			DH2 family			DH3 family			DH4 family			DH5 family					
	Sire	Dam	Progeny	Total loci	Dam allele	Sire* allele	Total loci	Dam allele	Sire* allele	Total loci	Dam allele	Sire* allele	Total loci	Dam allele	Sire* allele	Total loci	Dam allele	Sire* allele
Marker type I	??	aa	aa	149	121	28	141	123	18	166	127	39	155	130	25	160	146	14
Marker type II	??	bb	bb	81	56	25	73	57	16	75	53	22	78	60	18	73	63	10
Marker type III	??	cc	cc	7	5	2	4	3	1	8	6	2	6	6	0	7	6	1
Total				237	182	55	218	183	35	249	186	63	239	196	43	240	251	25

*: These are potential sire alleles or non-maternal alleles.

Table 5.10: The summary table of informative markers in haploid family (non-duplicated loci)

Haploid Fam	Sire	Dam	Progeny	Total loci	Dam allele	Sire allele
1. 100% informative markers	aa	bb	bb	8	2	6
2. 100% informative markers	ab	cc	cc	1	0	1
3. Less informative markers	aa	ab	aa, bb	15	0	1
4. Less informative markers	ab	aa	aa	43	0	21
Total				67	2	29

5.4.7 Microsatellites

A total of 363 (11 loci x 33 samples) genotypes were analysed using 3 multiplex sets (see Table 5.11) from 33 individuals (including outbred founders to G1 fish and the subsequent putative clonal fish-G2 families). However, only 9 out of 11 microsatellite loci successfully amplified and therefore could be used for investigation of homozygosity in both G1 and G2 families. One locus (0177) in MP-3 set was found to be out of Hardy-Weinberg equilibrium. However given the limited amount of individuals (five fish) in each family this was ignored. Similarly another two loci (8828 and 5488) represented bias towards fixation of one allele in the progeny; this was also ignored within the scope of the study. Thus, no loci were removed from the dataset for such reasons. In total, 3 loci (8592, 5794 and 8302) out of 9 were monomorphic (homozygous) in the dam, such loci were still informative to detect any potential sire contribution or an allele that does not corresponds to dam's genotype.

For all 9 microsatellite loci, there was no sire allele detected either in G1 or in G2 families. All 9 microsatellite loci were homozygous for the maternally derived allele among the progeny of both families (Table 5.11). Microsatellite results were in accordance with SNP data in G1 family but not among the G2 families (see Discussion).

The physical position of all microsatellite loci used in the present study was identified on genebank (NCBI) and visualised on a physical map alongside with 18 Norwegian microsatellites (Figs 5.3 and 5.4). In total 11 microsatellite loci used in the study were located in 10 different chromosomes (one locus, 5794, represented identical matches in two different chromosomes located in Ssal_02 and Ssal_05, named as 5794a and 5794b, respectively). In total of 18 Norwegian microsatellites, 11 different chromosomes was

identified in the Atlantic salmon genome, while 2 loci, SSspG7 and SSss1605 (Paterson et al., 2004) were identified in unassigned parts of the genome assembly, and one locus, MCHI (Grimholt et al. 2002) gave multiple hits for chromosome Ssal_27. The details regarding to position of microsatellites is provided in Table S2, (available online). However due to the lack of gene-centromere map in Atlantic salmon, centromeres could not be located in physical maps.

Table 5.11: Inheritance of microsatellite alleles (in bp) from outbred founders to G1 and G2 families. (--/-- represents missing data).

	Sample Information	Multiplex set_1			Multiplex set_2			Multiplex set_3		
		CAO53480	BG935488	CAO38592	CAO48828	CAO51136	CAO55301	CAO60177	CB515794	CAO48302
	1.Outbred dam	273/294	238/246	393/393	276/278	340/370	230/260	340/348	307/307	233/233
	2.Outbred sire	276/280	198/238	383/383	270/276	340/367	254/254	328/360	313/313	268/268
	3.G1_DH91	273/273	246/246	393/393	278/278	370/370	230/230	348/348	307/307	233/233
	4.G1_DH93	294/294	238/238	393/393	278/278	370/370	260/260	348/348	307/307	233/233
	5.G1_DH115	294/294	246/246	393/393	278/278	370/370	260/260	348/348	307/307	233/233
	6.G1_DH133	273/273	246/246	393/393	278/278	340/340	260/260	348/348	307/307	233/233
	7.G1_DH154	273/273	246/246	393/393	276/276	340/340	230/230	348/348	307/307	233/233
	8.G1_DH224	273/273	246/246	393/393	278/278	340/340	230/230	348/348	307/307	--/--
D H 1	9. G2_Clone2.1	273/273	246/246	393/393	278/278	370/370	230/230	348/348	307/307	--/--
	10.G2_Clone 2.2	273/273	246/246	393/393	278/278	370/370	230/230	348/348	307/307	--/--
	11.G2_Clone 2.3	273/273	246/246	393/393	278/278	370/370	230/230	348/348	307/307	--/--
	12.G2_Clone 2.4	273/273	246/246	393/393	278/278	370/370	230/230	348/348	307/307	233/233
	13.G2_Clone 2.5	273/273	246/246	393/393	278/278	370/370	230/230	348/348	307/307	233/233
D H 2	14.G2_Clone 3.1	294/294	238/238	393/393	278/278	370/370	260/260	348/348	307/307	233/233
	15.G2_Clone 3.2	294/294	238/238	393/393	278/278	370/370	260/260	348/348	307/307	233/233
	16.G2_Clone 3.3	294/294	238/238	393/393	278/278	370/370	260/260	348/348	307/307	233/233
	17.G2_Clone 3.4	294/294	238/238	393/393	278/278	370/370	260/260	348/348	307/307	233/233
	18.G2_Clone 3.5	294/294	238/238	393/393	278/278	370/370	260/260	348/348	307/307	233/233
D H 3	19.G2_Clone 4.1	273/273	246/246	393/393	278/278	340/340	260/260	348/348	307/307	233/233
	20.G2_Clone 4.2	273/273	246/246	393/393	278/278	340/340	260/260	348/348	307/307	233/233
	21.G2_Clone 4.3	273/273	246/246	393/393	278/278	340/340	260/260	348/348	307/307	233/233
	22.G2_Clone 4.4	273/273	246/246	393/393	278/278	340/340	260/260	348/348	307/307	233/233
	23.G2_Clone 4.5	273/273	246/246	393/393	278/278	340/340	260/260	348/348	307/307	233/233
D H 4	24.G2_Clone 5.1	273/273	246/246	393/393	276/276	340/340	230/230	348/348	307/307	233/233
	25.G2_Clone 5.2	273/273	246/246	393/393	276/276	340/340	230/230	348/348	307/307	233/233
	26.G2_Clone 5.3	273/273	246/246	393/393	276/276	340/340	230/230	348/348	307/307	233/233
	27.G2_Clone 5.4	273/273	246/246	393/393	276/276	340/340	230/230	348/348	307/307	233/233
	28.G2_Clone 5.5	273/273	246/246	--/--	276/276	340/340	230/230	348/348	307/307	233/233
D H 5	29.G2_Clone 1.1	273/273	246/246	393/393	--/--	340/340	230/230	348/348	307/307	233/233
	30.G2_Clone 1.2	273/273	246/246	393/393	278/278	340/340	230/230	348/348	307/307	233/233
	31.G2_Clone 1.3	273/273	246/246	393/393	278/278	340/340	230/230	348/348	307/307	233/233
	32.G2_Clone 1.4	273/273	246/246	393/393	278/278	340/340	--/--	348/348	307/307	233/233
	33.G2_Clone 1.5	273/273	246/246	393/393	278/278	340/340	230/230	348/348	307/307	233/233

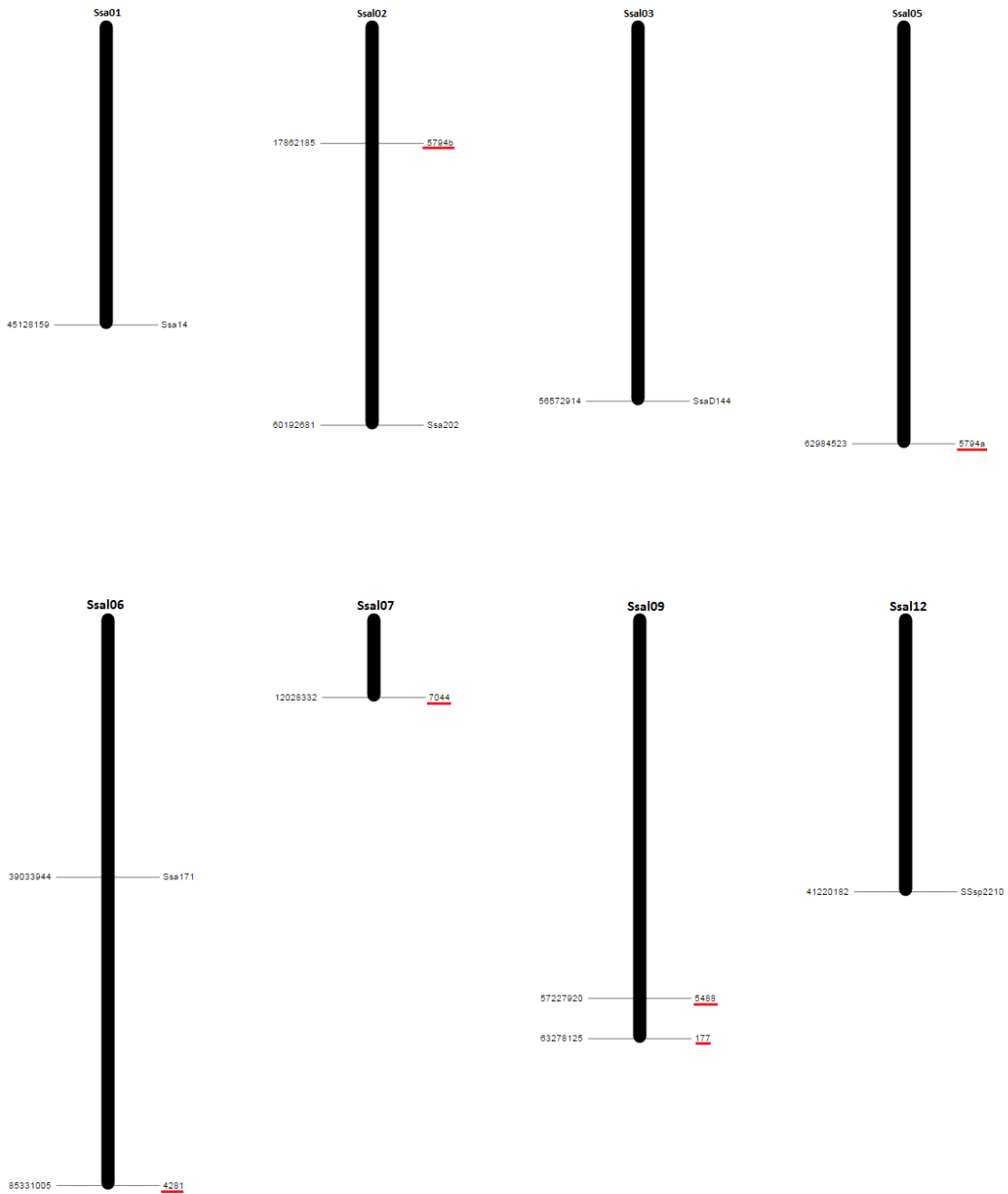


Figure 5.3: Physical position of microsatellites markers used in the present study alongside with Norwegian microsatellites. Red underlined loci represent the IOA microsatellites (continued in the next Figure). The length of the chromosomes is in accordance with the karyotype of Atlantic salmon.

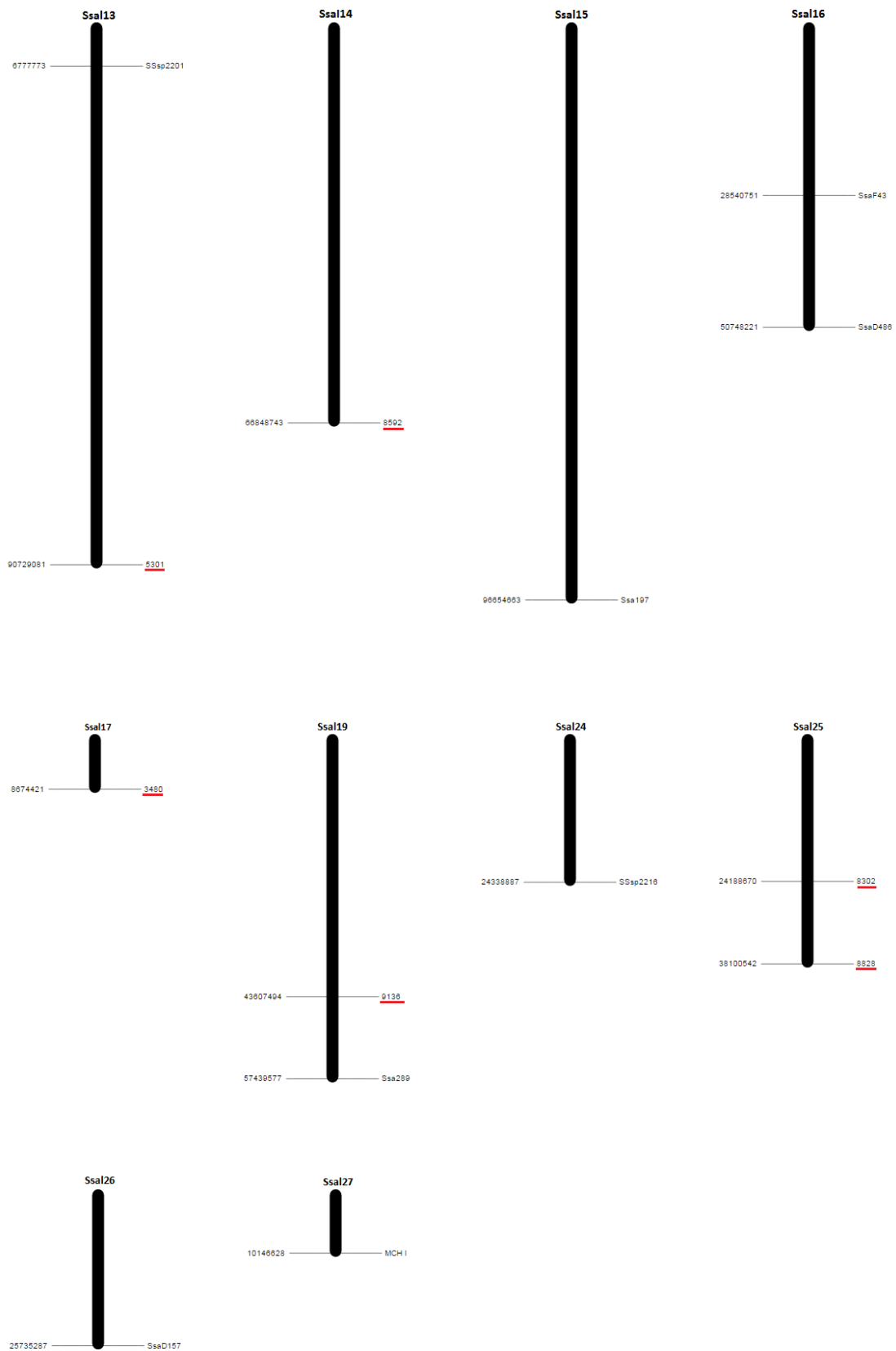


Figure 5.4: Physical position of microsatellites markers used in the present study alongside with Norwegian microsatellites. Red underlined loci represent the IOA microsatellites. The length of the chromosomes is in accordance with the karyotype of Atlantic salmon.

5.5 Discussion

In this study, we screened 6 G1 fish (clone founders) propagated from a single outbred female and 5 G2 fish per family each coming from the G1 fish (one of the G1 fish failed to have progeny) for isogenicity and we demonstrated all 6 G1 fish had completely homozygous genomes as opposed to varying levels of residual sire contribution observed in all 5 G2 families. This was a pilot study to investigate for genome-wide homozygosity in putative isogenic clonal lines using a novel approach for a species with a duplicated genome, this study also had a fundamental aim of verifying an optimised sperm UV-irradiation protocol in the Atlantic salmon through the use of high-throughput sequencing platforms compared to the few tens of microsatellites used in the past.

The present research demonstrated that homozygous, doubled haploid Atlantic salmon can be produced through pressure shock treatment of eggs activated by sperm treated with ultraviolet light. Aside from reduced representation of entire genome of salmon used in the present study, most studies have utilised relatively few loci to verify homozygosity in potential mitotic gynogenetic or androgenetic clone founders. Sarder et al. (1999) used the *ADA* locus and DNA fingerprinting to verify homozygosity in both G1 and G2 in Nile tilapia (*Oreochromis niloticus*), Galbusera et al. (2000) used 5 microsatellite loci in African catfish (*Clarias gariepinus*) while Bertotto et al. (2005) used 6 microsatellite loci in European seabass (*Dicentrarchus labrax*). Ezaz et al. (2004) used 12 AFLP markers and 7 microsatellite markers to verify the isogenic status of clonal lines in Nile tilapia (*O. niloticus*) while Ottera et al. (2011) used 5 microsatellite loci to verify the lack of paternal contribution in meiotic gynogenetic Atlantic cod (*Gadus morhua*). A recent study by Hou et al. (2015) used 30 microsatellite loci to determine the homozygosity of androgenetic progeny in zebrafish

(*Danio rerio*) which had been produced by applying a cold shock to fertilised eggs as a novel method for generating clonal lines by replacing hazardous UV, gamma, X-ray irradiation.

While the above studies represent a portion of the verification studies carried out on isogenic fish lines chronologically, it also highlights that fragmentary paternal contribution (maternal contribution in androgenetics) could have been missed due to the limited number of markers and marker technology used in the past. Here we used a method that applies reduced representation of the entire genome in the Atlantic salmon. Given the genome size of salmon (3.4×10^9 bp) is as big as the human genome, more markers were needed to ensure capturing any partial paternal fragment. To test the genome-wide coverage of the SNP markers generated by ddRADseq in the present study, we used a recent high density genetic linkage map (Gonen et al., 2014b) of the Atlantic salmon, RAD library constructed by using *SbfI* RE. Then, we created a local nucleotide database from the markers successfully assigned to the linkage map and Blasted all polymorphic ddRAD loci against them on BioEdit v7.2.5 (Hall, 1999). Results showed our markers were a subset (15%) of Gonen et al. (2014) and our markers were represented in all linkage groups ranging from 1 to 7% which proves the genome-wide coverage of markers in the existing study. Likewise, the distribution of markers within the linkage groups in both studies were similar to one another.

As both G1 and G2 families were produced by applying mitotic and meiotic gynogenesis respectively, the resultant progeny from both families were expected to be all female and be homozygous at all loci. So we expected the vast majority of the ddRAD loci to be homozygous with few homologous loci that might show PSVs or MSVs as a result of the WGD in the salmonids. This was proven to be true in G1 family

(Table 5.8) after removal of multi-copy loci that complicated the verification study. However, the sire contribution was consistent in all 5 of the progeny of G2 families even after removing of multi-copy loci (see Table 5.9). This was due to the existence of partial sire contribution in the G2 families, which was confirmed by investigation of one-copy loci following BLAST analysis (see Table 5.9). The sires used to propagate the G2 families came from a pool of 2 males, no DNA was provided for either of the males. The fact that sire alleles were seen consistently among the progeny in each G2 families provides solid evidence that UV irradiation of milt step was sub-optimal. This can be broken down into two scenarios as follows: (i) technical and/or (ii) logistic reasons. The UV lamp used might have degraded since being used to produce G1 clone founders or it can be as simple as sperm dilutions were not applied to both males individually hence might have resulted in non-optimal concentration, shading of milt for the UV irradiation. Regardless of what is the cause for non-optimal UV dose, this research still holds the promise for reproducing clone founders (G1) in the next spawning season to establish isogenic clonal populations. Similar trend was seen in haploid family where one-copy loci did not rule out the existence of sire alleles.

Recent advances achieved in genotyping using NGS platforms make it possible to work with any organism without requiring prior genetic information. (e.g: SNP markers are detected following ddRADseq protocol as opposed to knowing the sequence information of up- and down-stream primer binding sites for microsatellites). Therefore, SNP identification through the use of high throughput sequencing is a straightforward process in diploid genomes where only allelic forms of the polymorphism are detected and alignment of sequences are screened to identify different alleles at the same locus. However, in duplicated genomes this process is more complex due to the occurrence of three types of sequence variants (PSVs, MSVs and SNPs) and all appear to be

polymorphic. In the earliest attempts at identifying SNP markers, (Smith et al., 2005) removed 32% of the data as being duplicated loci in salmonids through Sanger sequencing of 89kb. Later, as the NGS became more available; researchers used excessive heterozygosity as a way of filtering the dataset to remove the fixed sites, PSVs, from the actual SNP panel in duplicated genomes (Gonen et al., 2014; Hohenlohe et al., 2011). Yet, none of the strategies guarantees the successful removal of all PSVs and MSVs till better tools become available (Gidskehaug et al., 2010).

To deal with the difficulties encountered in duplicated genomes, we used Blast analysis in order to identify copy number of given loci using all available versions of the genome assembly on genebank (NCBI). The newer versions of the assembly (*Ssal_v2* and *Ssal_v4*) dramatically increased the size of the assembly. The number of duplicates also increased because the initial assembly was missing in such regions. Nonetheless, it must be acknowledged that the present research was heavily depended on available genomic resources in the Atlantic salmon as well as the bioinformatics tools developed. None of the potential paternal contributor loci would have been identified properly as being homologs or repetitive elements if the genome assembly for the Atlantic salmon was not available. The comparison between initial assembly *Ssal_v1* and the current assembly *Ssal_v4* clearly shows that, a good quality reference assembly is an essential while working with duplicated genomes (Table 5.5).

Microsatellite genotyping carried out in the present research was used as a means to double check the existence of sire alleles in both G1 and G2 families after observing high levels of heterozygosity among the G2 families. This was not initially included in the scope of the project. Inheritance of microsatellite alleles at nine loci clearly showed exclusive transmission of maternal genome in both G1 and G2 families. This was also

in accordance with 18 microsatellite markers that had previously been used to genotype the same samples and suggested homozygosity of all samples (Glover et al., 2009). However, ddRADseq data (one-copy polymorphic loci) confirmed the homozygosity of the G1 progenies while detecting varying levels of non-maternal alleles in G2 families (Table 5.8 and 5.9). This can be explained in two scenarios: i. the position of microsatellite markers can be in the homozygosity block of chromosomes, and/or closer to centromeric regions or ii. mutations that occur in the primer binding sites might cause the allele not to be seen in the G2 families where sire allele previously detected, a heterozygote may be seen as a homozygote; these missing cases are referred to as “null” alleles (Chapuis & Estoup, 2007). Although the physical position of nine microsatellite loci used in the present study was identified diversely covering 18 out of 27 chromosomes in the genome of Atlantic salmon, it was not possible to detect the position of markers relative to centromere and/or distal regions where higher recombination is more likely to take place (Figs 5.3 and 5.4; Table S2). Thus one can only speculate on the position of microsatellite markers as being physically closer to potential centromeric regions. Considering the marked difference in crossover frequencies between sexes in Salmonidae family members (Allendorf et al., 2015 refs cited therein; Gharbi et al., 2006; Gonen et al., 2014), it is difficult to predict the discriminatory power of microsatellite markers based on available genetic resources as of yet. There were deviations observed in some of the microsatellite loci (e.g: 0177, 8828, 5488) from the expected Mendelian ratio. These deviant segregations of alleles are possibly results of reduced viability of homozygous individuals, since there is no heterozygotes to unmask any recessive lethal or semi-lethal gene that is tightly linked to a specific locus, such deviations can be seen in favour of expression of survival genotypes (Lahrech et al., 2007). Given the small number of individuals screened for

homozygosity in both G1 (6 fish) and G2 (5 fish in each family, 25 in total), this was acceptable. The results of ddRADseq data, however, showed a successful elimination of any sire contribution in the G1 progeny. This, also, indicated that next generation sequencing technologies can be used as a means to detect any potential residual parental chromosome fragments in species of duplicated genomes as in Atlantic salmon; this depending on size of the fragments and the coverage of loci.

Stack is an open source pipeline, which is designed initially for *de-novo* assemblies to make it more flexible for organisms with or without the availability of reference genome assemblies. One benefit of the pipeline is its modular structure, which allows Stacks to be used for numerous scenarios, and mapping approaches. This, however, also brings along an increased possibility of erroneous results if not handled by an expert bioinformatician. One example would be specifying `-m` parameter; minimum number of identical raw reads required to create a stack (later to be used for genotype calls by matching loci/stacks with parents, in the case of family analysis). However, setting this parameter up to a certain value avoids getting the best stacks by biasing them towards homozygotes. For example: using `-m:6` for an individual genotype call means that samples with a read depth of 6 can never be a heterozygote, because an alternate allele will have at least 1 read thus the main alleles with 5 reads will not be output by the pipeline. This leads to overlooking an alternate allele and filtering out SNPs based on major stacks parameters. Allele depth of 5:1, 4:2, 3:3 (gives a total of 6 individual genotype coverage) will be scored as homozygote although alternate alleles ensure clear heterozygosity case with no sequencing error. The only case of heterozygosity call appears having read depth over set up `-m` parameter as a minimum of 6:1 in the allele depth of 7 for given individual genotype and onward of allele depths. (e.g: 7 read depth, can call heterozygote in 6:1, 8 read depth can call heterozygote in 7:1, 6:2, 9 read depth

can call heterozygotes in 8:1, 7:2, 6:3, the rest of the alleles depth 5:2, 4:3; 5:3, 4:4; 5:4 will be called as homozygote as these do not have an at least an allele with 6 reads, respectively in the individual genotype coverage of 7, 8 and 9). In the present verification study such parameters were closely monitored and kept identical in G1 and G2 families. However, Stacks parameters needed to be re-set after *process-radtag* module during the analysis of haploid family. Since female parent (MO866; 135,433) produced almost four times less reads compared to male parent (MO865; 517,406) specifying a minimum number of identical raw reads (-m) required to create a stack decreased from 10 (initial analysis) to 4 in the haploid family. This significantly improved most of female genotype missed loci to be genotyped accurately but in less depth compared to progeny and the male parent (see Tables 5.3 and 5.10). Yet, this modification did not rule out the heterozygotes observed in haploid family, caused by the sub-optimal UV doses. The second example of using Stacks pipeline carefully can be explained with the existence of secondary reads during SNP calling. Although these are referred to as sequencing errors most times, some are playing a significant role during the SNP calling step of the pipeline. As the minimum stack depth (-m) controls the number of raw reads required to form initial stacks, these are also termed primary reads identical to one another. If the depth of coverage for any particular stack is below this threshold, then the allele will not be formed and these will be called secondary reads and will be temporarily set aside by the algorithm. However, secondary reads are incorporated into the analysis once the loci are formed (by default). This process provides more depth to the SNP calling model for detecting polymorphisms for an increasing likelihood estimate. One aspect that cannot be stressed enough is that each of these secondary reads has a single, best-alignment to an existing locus to be incorporated into a single locus (not to multiple loci). For example in the case of

applying high values for the (-m) parameter many alleles might be missed. However, later with the incorporation of secondary reads alleles with coverage lower than (-m) value can be rescued. This default setting to secondary reads is designed to give a better likelihood estimate and aides the SNP calling model in detecting polymorphisms. This however, constitutes what needs to be eliminated for verification studies due to possibility of carrying sequencing errors.

5.5.1 Conclusion

In an effort to verify genome-wide isogeny to establish an optimised sperm UV irradiation protocol for Atlantic salmon, the present study used the power of high-throughput sequencing technology, starting from outbred founders to first generation clone founders (G1) and to the putative isogenic clonal lines (G2) in duplicated genome of salmon. This work clearly demonstrated that the first generation of clone founders represented genome-wide homozygosity while putative clones represented varying levels of residual non-maternal (paternal) fragments (10-25%) shared among all families, probably because of a sub-optimal UV irradiation during the propagation of second generation. To our best knowledge this study is the first verification work undertaken through the use of NGS platforms to verify large-scale homozygosity in the Atlantic salmon genome as opposed to handful of microsatellite markers used in the past. Considering the existing complications of Atlantic salmon genome post-WGD, the approach used here provided evidence on NGS technologies can be used as an improved way of detecting any contribution from the inactivated parental genome. This applies even in the ancestrally tetraploid genomes such as that of Atlantic salmon with an exception of having a good quality reference assembly available. On the basis of having verified homozygous G1 fish with an optimised UV irradiation protocol, the future research will involve gynogenetic reproduction of successfully produced

homozygous clone founders (G1) in the present research to establish isogenic clonal populations useful for identification and estimation of genetic and environmental components of trait variation, particular interest for aquaculture related research (Bongers et al., 1997b; Tanck et al., 2002; Zimmerman et al., 2004).

Acknowledgement

Authors are grateful to Jose C.M.V.G (Landcatch, UK) for the supply of the parents of haploid family.

Chapter 6

Restriction digestion inhibition observed in the early developmental stages of Nile tilapia (*Oreochromis niloticus*)

Abstract

The intention of the present study was to use an experimentally produced meiotic gynogenetic family of Nile tilapia (i) to explore potential residual genome-wide paternal genetic contribution and (ii) to construct a SNP-based genetic linkage map using double-digest restriction associated DNA sequencing. However, due to an unexpected inhibitory mechanism observed in the DNA of *O. niloticus* larvae, restriction digestion could not be efficiently achieved. These results were confirmed in a full-sib family produced and sampled at different developmental stages. However, the inhibition mechanism in Nile tilapia larvae could not be identified.

Keywords: *Isogenic clonal lines, meiotic gynogenetics, ddRADseq, Nile Tilapia, aquaculture*

6.1 Introduction

Gynogenesis is a form of uniparental reproduction, which leads into progeny with only maternal genome contribution. However problems associated into induction protocols are commonly observed in gynogenesis either due to sub-optimal UV irradiation of sperm (since 100% maternal genome transmission is desired) or spontaneous arisal of meiotic gynogenetics among the doubled haploid mitotic gynogenetic progeny. As meiotic gynogenetics results from capturing of the second polar body, these individuals carry some heterozygosity from the results of female crossover. Thus, their presence may interfere with the reliable production of isogenic clonal lines in the second generation. Therefore, meiotic gynogenetics (with varying level of heterozygosity) need to be detected and separated from the completely homozygous doubled haploid G1 progeny before proceeding into the second generation.

A large number of markers is required for both purposes. Informative genetic markers, capable of detecting both potential contribution from irradiated gametes and discriminating mitotic and meiotic gynogenetics, are required to verify the production of the initial doubled haploids (G1), and the development of isogenic lines from these. A variety of markers have been used for this purpose (including pigmentation genes, allozymes, DNA fingerprinting, microsatellites, etc: Komen and Thorgaard, 2007). One important aspect is the number of available markers (e.g. large numbers are desirable if potential inheritance of chromosome fragments from irradiated gametes is to be detected). Considering that most teleosts have around 22-25 each chromosome pairs, needs to be represented with a decent number of markers. The second aspect to take into account is the ability of markers to discriminate between mitotic and meiotic gynogenetics: those that are located close to centromeric regions will be compromised with respect to their ability to detect crossover events. This is a key requirement when differentiating between mitotic and meiotic gynogens; i.e. informative telomeric markers will be heterozygous in meiotic gynogenetics and homozygous in mitotic gynogenetics, while centromeric markers will largely be homozygous in both types, thus lacking any discriminatory power (Danzmann & Gharbi, 2001).

Research advances in isogenic clonal lines have become even more significant as the power of isogenic clonal lines approach has been combined with recent advances in next generation sequencing technologies. Such technologies have the capacity to discover thousands of SNP markers simultaneous per individual at decreasing costs. This provides a unique opportunity to more accurately assess the effectiveness of both meiotic and mitotic gynogenetics as a means of reliable production of isogenic clonal fish lines. In this study SNP markers that were generated by using ddRADseq (Peterson et al., 2012) were initially planned to be used to explore genome-wide paternal genetic contribution in

an experimentally produced meiotic gynogenetic family of Nile tilapia (*Oreochromis niloticus*) and develop a SNP-based genetic linkage map was based on the meiotic gynogenetic Nile tilapia family. A few haploids from the same family were also used as an additional control for UV irradiation treatment while a bi-parental control family from the same parental source was additionally included in the library so as to estimate the percentage of markers that showed non-Mendelian inheritance. This is more likely to be a potential problem in big datasets with large set of markers. Typically less than 5% is in the acceptable limits.

6.2 Materials and Methods

6.2.1 Ethics statement

All working procedures used in the chapter complied with the United Kingdom Animals (Scientific Procedures) Act 1986 and were approved by the ethics committee of the University of Stirling.

6.2.2 Experimental design to sampling

This study was carried out using a single female (blonde type, PIT: 00-0690-A589) and a male (wild type, PIT: 00-068C-DA03). A ready to spawn female with a swollen abdomen and urogenital papilla, presenting pre-spawning behaviour such as nest building and cleaning the bottom of the tank, was used for stripping eggs of good quality (See Chapter 2 for the details of the procedures applied including collection of gametes, UV irradiation of sperm, fertilisation, heat shock to suppress exclusion of second polar body, egg incubation, larval development and sampling). The pigmentation character of wild type male and haploid group provided controls for UV irradiation of sperm (e.g: developing

embryos would present pigmentation in the case of residual sire contribution). Haploids cannot survive beyond hatching and they represent the so-called haploid syndrome characteristic, with smaller embryos.

Table 6.1: Schematic diagram of the experimental design

Groups	Type	Description																
Bi-parental control	2n-Control	<ul style="list-style-type: none"> • Ordinary fertilisation 																
Haploid	n-Control	<ul style="list-style-type: none"> • Activation via UV-irradiated sperm • No heat shock 																
Meiotic gynogenetic	2n-Meiotic gynogenetic	<ul style="list-style-type: none"> • Activation via UV-irradiated sperm • Heat shock applied 																
		<table border="1"> <thead> <tr> <th>Groups</th> <th>Shock Started (mins)</th> <th>Shock Ended (mins)</th> <th>Temp. (°C)</th> </tr> </thead> <tbody> <tr> <td>HS-1</td> <td>4'45''</td> <td>8'15''</td> <td>41.5±0.5</td> </tr> <tr> <td>HS-2</td> <td>5'00''</td> <td>8'30''</td> <td>41.5±0.5</td> </tr> <tr> <td>HS-3</td> <td>5'15''</td> <td>8'45''</td> <td>41.5±0.5</td> </tr> </tbody> </table>	Groups	Shock Started (mins)	Shock Ended (mins)	Temp. (°C)	HS-1	4'45''	8'15''	41.5±0.5	HS-2	5'00''	8'30''	41.5±0.5	HS-3	5'15''	8'45''	41.5±0.5
		Groups	Shock Started (mins)	Shock Ended (mins)	Temp. (°C)													
		HS-1	4'45''	8'15''	41.5±0.5													
HS-2	5'00''	8'30''	41.5±0.5															
HS-3	5'15''	8'45''	41.5±0.5															

In total three experimental meiotic gynogenetic groups were produced from a single parental set depending on heat shock start time (4'45'', 5'00'' and 5'15'' AF, for meiotic gynogenetic heat shock groups 1, 2 and 3 respectively, see Table 6.1). Each experimental group received the same duration of heat shock for 3'30'' minutes. The UV irradiation dose and the heat shocks employed were according to the protocol of Sarder et al. (1999). The bi-parental and the haploid group were also produced alongside the experimental groups. Each group was placed into a separate egg incubator until sampling was carried out which was based on the expected survival time of the haploid group. Since haploids do not survive long beyond hatching, sampling was carried out at hatching, roughly 4 days AF. Each group was placed into incubation water in a plastic petri dish to first observe the developmental stage under dissection microscope then 1-2 drops of benzocaine stock solution was added into petri dishes prior to sampling of each larva into a separate tube containing 1 mL of absolute EtOH. Table 6.2 lists the samples used for constructing ddRAD library.

Table 6.2: List of the samples used.

Sample ID	Type	Sample ID	Type
MO-446	Sire (D8a: 00-068C-DA03)	MO-378	Meiotic gynogenetic (HS-2)
MO-447	Dam (C2a: 00-069O-5A89)	MO-379	Meiotic gynogenetic (HS-2)
MO-262	Bi-parental Control	MO-380	Meiotic gynogenetic (HS-2)
MO-263	Bi-parental Control	MO-382	Meiotic gynogenetic (HS-2)
MO-264	Bi-parental Control	MO-383	Meiotic gynogenetic (HS-2)
MO-265	Bi-parental Control	MO-384	Meiotic gynogenetic (HS-2)
MO-266	Bi-parental Control	MO-385	Meiotic gynogenetic (HS-2)
MO-267	Bi-parental Control	MO-386	Meiotic gynogenetic (HS-2)
MO-268	Bi-parental Control	MO-387	Meiotic gynogenetic (HS-2)
MO-269	Bi-parental Control	MO-388	Meiotic gynogenetic (HS-2)
MO-270	Bi-parental Control	MO-389	Meiotic gynogenetic (HS-2)
MO-271	Bi-parental Control	MO-390	Meiotic gynogenetic (HS-2)
MO-342	Haploid	MO-391	Meiotic gynogenetic (HS-2)
MO-343	Haploid	MO-392	Meiotic gynogenetic (HS-2)
MO-344	Haploid	MO-393	Meiotic gynogenetic (HS-2)
MO-345	Haploid	MO-394	Meiotic gynogenetic (HS-2)
MO-346	Haploid	MO-395	Meiotic gynogenetic (HS-2)
MO-348	Haploid	MO-396	Meiotic gynogenetic (HS-3)
MO-349	Haploid	MO-397	Meiotic gynogenetic (HS-3)
MO-350	Haploid	MO-398	Meiotic gynogenetic (HS-3)
MO-351	Haploid	MO-399	Meiotic gynogenetic (HS-3)
MO-352	Meiotic gynogenetic (HS-1)	MO-400	Meiotic gynogenetic (HS-3)
MO-353	Meiotic gynogenetic (HS-1)	MO-402	Meiotic gynogenetic (HS-3)
MO-354	Meiotic gynogenetic (HS-1)	MO-403	Meiotic gynogenetic (HS-3)
MO-355	Meiotic gynogenetic (HS-1)	MO-404	Meiotic gynogenetic (HS-3)
MO-356	Meiotic gynogenetic (HS-1)	MO-405	Meiotic gynogenetic (HS-3)
MO-357	Meiotic gynogenetic (HS-1)	MO-406	Meiotic gynogenetic (HS-3)
MO-358	Meiotic gynogenetic (HS-1)	MO-407	Meiotic gynogenetic (HS-3)
MO-359	Meiotic gynogenetic (HS-1)	MO-408	Meiotic gynogenetic (HS-3)
MO-360	Meiotic gynogenetic (HS-1)	MO-409	Meiotic gynogenetic (HS-3)
MO-361	Meiotic gynogenetic (HS-1)	MO-410	Meiotic gynogenetic (HS-3)
MO-362	Meiotic gynogenetic (HS-1)	MO-411	Meiotic gynogenetic (HS-3)
MO-363	Meiotic gynogenetic (HS-1)	MO-412	Meiotic gynogenetic (HS-3)
MO-364	Meiotic gynogenetic (HS-1)	MO-413	Meiotic gynogenetic (HS-3)
MO-365	Meiotic gynogenetic (HS-1)	MO-414	Meiotic gynogenetic (HS-3)
MO-366	Meiotic gynogenetic (HS-1)	MO-415	Meiotic gynogenetic (HS-3)
MO-367	Meiotic gynogenetic (HS-1)		
MO-368	Meiotic gynogenetic (HS-1)		
MO-369	Meiotic gynogenetic (HS-1)		
MO-372	Meiotic gynogenetic (HS-1)		
MO-373	Meiotic gynogenetic (HS-1)		
MO-374	Meiotic gynogenetic (HS-2)		
MO-375	Meiotic gynogenetic (HS-2)		
MO-376	Meiotic gynogenetic (HS-2)		

6.2.3 DNA extraction and quantification

DNA was extracted from the whole larva (the yolk sac was removed where possible; this was easy in diploid groups (2n) but more difficult in the haploid (n) group where the yolk could not be separated from developing larvae) using the REALPure genomic DNA extraction kit (Real Laboratories, Spain) and treated with RNase to remove residual RNA from the samples. This kit also involved a protein precipitation step to remove degraded yolk proteins. Each sample was quantified by spectrophotometry (Nanodrop) and quality assessed by agarose gel electrophoresis, and was diluted to a concentration of 50 ng/ μ L in 5 mmol/L Tris, pH 8.5 as working solutions. A final, more accurate, fluorometric-based assessment of DNA concentration was then performed on all samples using the Qubit® dsDNA HS Assay Kit (Invitrogen, UK). Fluorescence measurements (20 μ L volumes) were performed on a 96 well qPCR thermal cycler (Quanta, Techne, UK), with Nile tilapia DNA concentrations being derived from a calibration curve generated from a set of standard dsDNAs. Based on these readings a final dilution of 10 ng/ μ L in 5 mM Tris, pH 8.5 was employed to be used at ddRAD library construction protocol of Nile tilapia samples.

6.2.4 ddRAD library preparation and sequencing

The ddRAD library preparation protocol used in the present study was initially based on Peterson et al. (2012) but slightly modified as described in detail elsewhere (Brown et al., 2016; Manousaki et al., 2015). However it is of prime importance to note that the ddRAD libraries used in this experiment were of similar to those described in the rest of the thesis where the modified procedures were verified. However each library differed from the others based on the number of PCR cycles used to enrich library, which was individually

set up based on the intensity of the library template on the agarose gel (see Chapter 2 for details).

The procedure described below explains the preparation of the second ddRAD library, which is the only one that was sequenced, while in total three ddRAD libraries were generated in this study. Out of the three libraries, the first and the second ddRAD libraries were constructed with the same samples as explained in 6.2.2 following the procedure explained below, while the third one was a control library, based on only DNA from four fin samples (6.3.3).

Briefly, a single restriction enzyme digestion / adapter ligation reaction was performed for each progeny sample, while triplicate reactions were made for both dam and sire DNA samples. This was employed to ensure high coverage in parental samples in order to more confidently assign true SNPs in the pedigree later. Each sample (21 ng DNA) was digested at 37°C for 90 minutes with 0.8 U *SbfI* ('rare' cutter, CCTGCA|GG motif) and 0.8 U *SphI* ('common' cutter, GCATG|C motif) high fidelity restriction enzymes (New England Biolabs; NEB) in a 6 µL reaction volume that included 1× CutSmart™ buffer (NEB). After cooling the reactions to room temperature, 3 µL of a premade barcode-adapter mix was added to the digested DNA, and incubated at room temperature for 10 min. This adapter mix comprised individual-specific barcoded combinations of P1 (*SbfI*-compatible) and P2 (*SphI*-compatible) adapters at 6 nM and 72 nM concentrations respectively, in 1× reaction buffer 2 (NEB). Adapters were compatible with Illumina sequencing chemistry (see Peterson et al. (2012) for details). The barcoded adapters were engineered such that adapter–genomic DNA ligations did not reconstitute RE sites, while residual RE activity limited concatemerization of genomic fragments. Ligation was performed over 195 min at 22°C by addition of a further 3 µL of a ligation mix

comprising 4 mM rATP (Promega, UK), and 2000 cohesive-end units of T4 ligase (NEB) in 1× CutSmart buffer.

The ligated samples were then combined into a single pool and were column-purified (MinElute PCR Purification Kit, Qiagen, UK). Size selection of fragments, c. 320 bp to 590 bp, was performed by agarose gel electrophoresis. Following gel purification (MinElute Gel Extraction Kit, Qiagen, UK) the eluted size-selected template DNA (64 µL in EB buffer) was PCR amplified (18 cycles PCR; 32 separate 12.5 µL reactions, each with 1.25 µL template DNA) using a high fidelity Taq polymerase (Q5 Hot Start High-Fidelity DNA Polymerase, NEB). The PCR reactions were combined (400 µL total), and column-purified (MinElute PCR Purification Kit). The 55 µL eluate, in EB buffer, was then subjected to a further size-selection clean up using an equal volume of AMPure magnetic beads (Perkin-Elmer, UK), to maximise removal of small fragments (less than ca. 200 bp). The final library was eluted in 20 µL EB buffer (ca.18.6 µL returned from paramagnetic bead-library mix) and sequenced over a full Illumina MiSeq runs (v2 chemistry, 300 cycle kit, 162 bp paired end reads; Illumina, Cambridge, UK; 9.5 pM library applied and both runs spiked over 2% Illumina phiX control DNA).

6.3 Results

6.3.1 The first ddRAD library

The double-digest restriction library was not sufficiently digested, as seen from the agarose gel (Fig 6.1A). Following test PCRs (Fig 6.1B), the amplification of the library was optimised with 1.5 µL template DNA in half reaction volume (12.5 µL) for 13 PCR cycles (Figure 6.1C) which was carried out in a total volume of 400 µL by splitting into 32 PCR tubes via bulk PCR. Quantification of template (0.124 ng/µL) and the final

library (2.58 ng/ μ L) revealed relatively poor yield with unusual size distribution towards bigger fragments (Fig 6.1C). A total of 18.6 μ L volume was returned from the AmPure beads clean-up, with a concentration of 2.58 ng/ μ L, available for sequencing run. This stock was stored in a freezer, however was not used for sequencing due to the unusual fragment size distribution (Fig 6.1C).

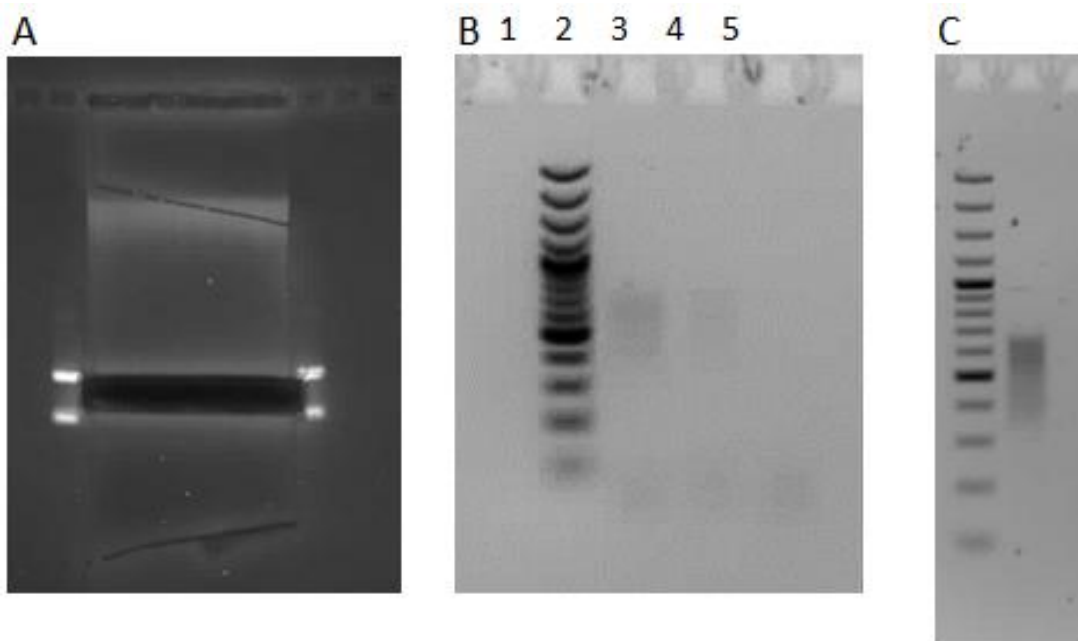


Figure 6.1: Gel images of the first ddRAD library constructed in Nile tilapia. (A) Library size selection gel loaded with 320-590 bp markers to detect the size of interest, (B) represents the results of test PCRs where template, 100 bp GeneRuler, amplicon of 16X PCR with 0.5 μ L template, amplicon of 13X PCR with 1 μ L template, and no template control (NTC) were loaded respectively from lanes 1 to 5 while (C) represents the final library run with 100 bp GeneRuler (note unusual size distribution evident with intense bigger fragments).

6.3.2 The second ddRAD library

Insufficient restriction digestion was observed in the second ddRAD library, similarly to the previous library constructed from the same samples in Nile tilapia (Fig 6.2A). Test PCRs were compared with that of previous library on the same gel (Fig 6.2B), revealing that the yield of the second ddRAD library was 2^5 times less than previous library (needed 5 PCR cycles more to reach the same yield). This was confirmed with 18X PCR

cycles for the enrichment of the library. A homogenous final library presented an even DNA smear in the expected size range distribution on the agarose gel (Fig 6.2C). Quantification of template (0.1 ng/μL) and the final library (7.81 ng/μL) showed that the total yield was higher in the second ddRAD library compared to first library (although based on more PCR cycles). A total of 18.6 μL library was returned from the AmPure bead clean-up and 2 μL of this was used for one sequencing run due to an even homogenous smear on the expected size range.

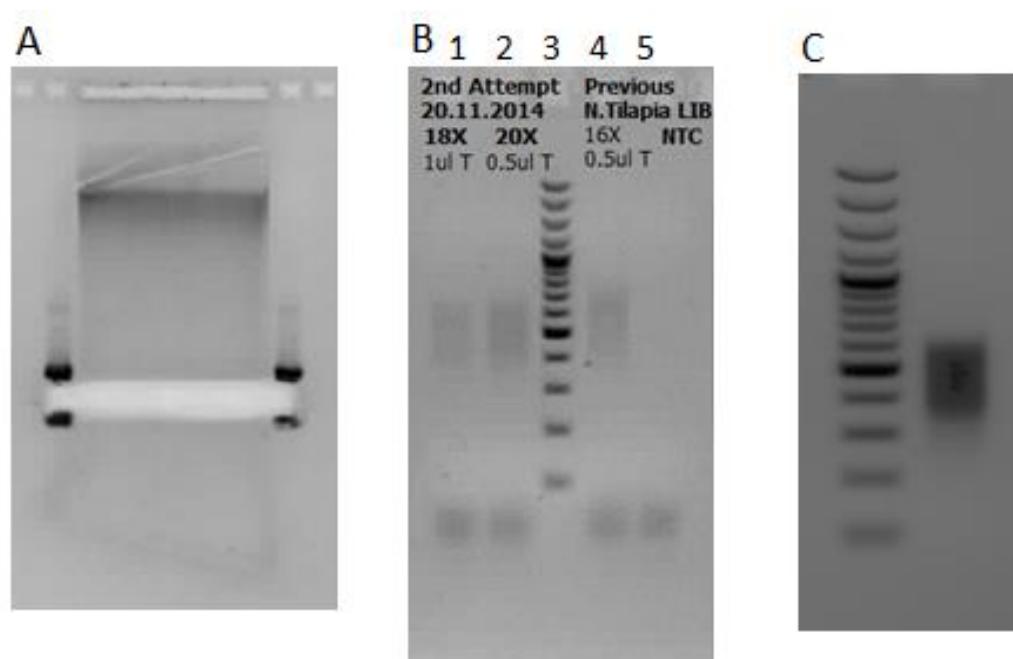


Figure 6.2: Gel images of the second ddRAD library in Nile tilapia. (A) Library size selection gel loaded with 320-590 bp markers to detect the size of interest, (B) represents the results of test PCRs where present and the previous ddRAD libraries constructed using the same samples in Nile tilapia were compared in terms of yields, lane 1 to 5 demonstrate respectively the results of amplicon of 18X PCR with 1 μL template and amplicon of 20X PCR with 0.5 μL template, 100 bp GeneRuler, amplicon of 16X PCR with 0.5 μL template from the previous ddRAD library in Nile tilapia and NTC while (C) represents the final library gel image with 100 bp GeneRuler.

Quality control of the raw reads carried out by FastQC v. 0.11.3 (Andrews, 2010) revealed a high quality sequencing run (Report S1 & S2). A total of 31,965,742 reads (each 162 bp long) were obtained (i.e: 15,982,871 paired-end) from the sequencing of the 87 individuals including parents, three HS groups of meiotic gynogenetics, haploids and

bi-parental controls (plus seven samples from another project). The Stacks pipeline (Catchen et al, 2011) was used for filtering of low quality reads (Phred33, quality score under 30; 233,008), ambiguous barcodes and ambiguous ddRAD tags (total 5,659,418) which left 26,073,316 high quality reads. Thus, 81.56 % of the raw reads were retained. However high variation was observed in terms of reads that were produced from the same samples (e.g: triplicated parental samples showed such variation clearly: see colour-coded parental sample cells in Fig 6.3. The majority of the samples failed to produce anywhere near the minimum threshold of 60,000 reads, thus no additional sequencing run was carried out to increase the depth of the coverage. Furthermore there was no correlation observed across the plate that could be explained by the use of a multi pipette.

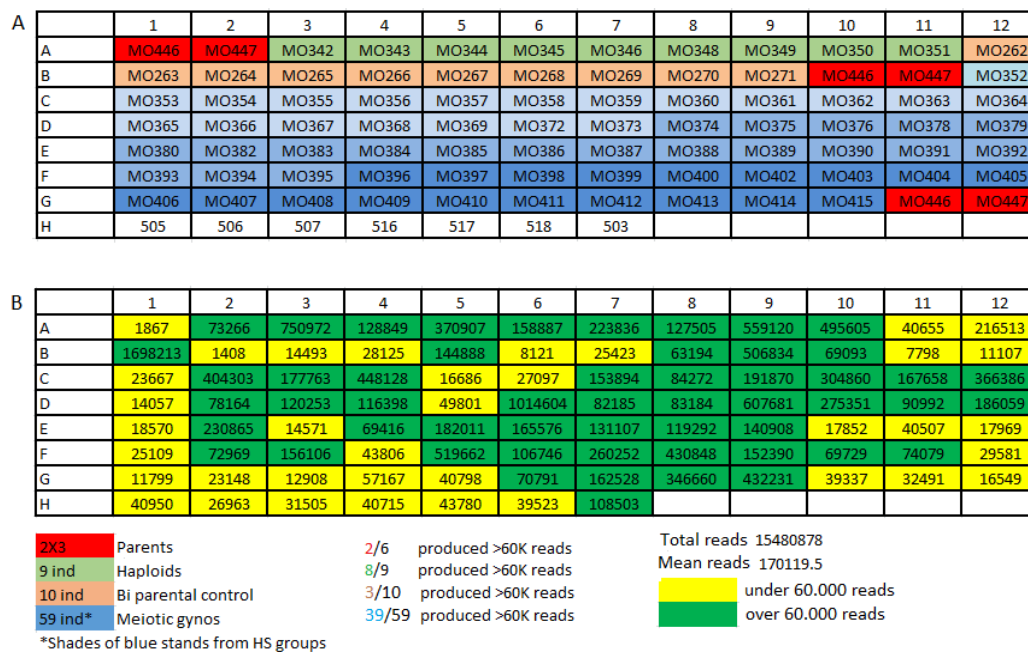


Figure 6.3: Diagram shows the results of one round of sequencing run on MiSeq in terms of reads that were produced. (A) shows the plate order of the samples that were used for the construction of the first and the second ddRAD library while (B) represents the filtered reads that were produced at the end of one MiSeq run. Conditional formatting option in excel was used for highlighting of reads under the threshold of 60,000 filtered reads per sample, yellow cells show reads detected under threshold while green cells show reads over threshold of 60,000 reads.

6.3.3 The third ddRAD library constructed as a control from fin clips

Sufficient double-digest restriction digestion with an intense smear on the agarose gel was observed in the control family (Fig 6.4A-note the intensity of the smear is much higher than both previous libraries produced in Nile tilapia involving larvae DNA samples, see Figs 6.1A and 6.2A respectively). Test PCRs revealed that optimisation of the library was ideal with 14 cycles using 1.5 μ L template DNA (Figure 6.4B). This library was used as a control to test whether or not consistent failure observed on construction of ddRAD library evident with variation detected in sequencing outcome was a human error during the wet-lab procedures.

The outcome of the control library proved an acceptable restriction digestion pattern from four parental fin clips that were replicated in the library (4 fins x 24 times = 96 well plate was used with the same concentrations of reagents, identical to previous library procedures). A multi pipette was used throughout the procedure (loading genomic DNA, master-mixes for both restriction digestion and ligation steps) therefore no variation was caused by pipetting. The number of PCR cycles required to amplify the final library was also within the desired limit (<16 cycles). Therefore, no further step was taken to amplify and/or finalise this control ddRAD library representing only parental fin samples.

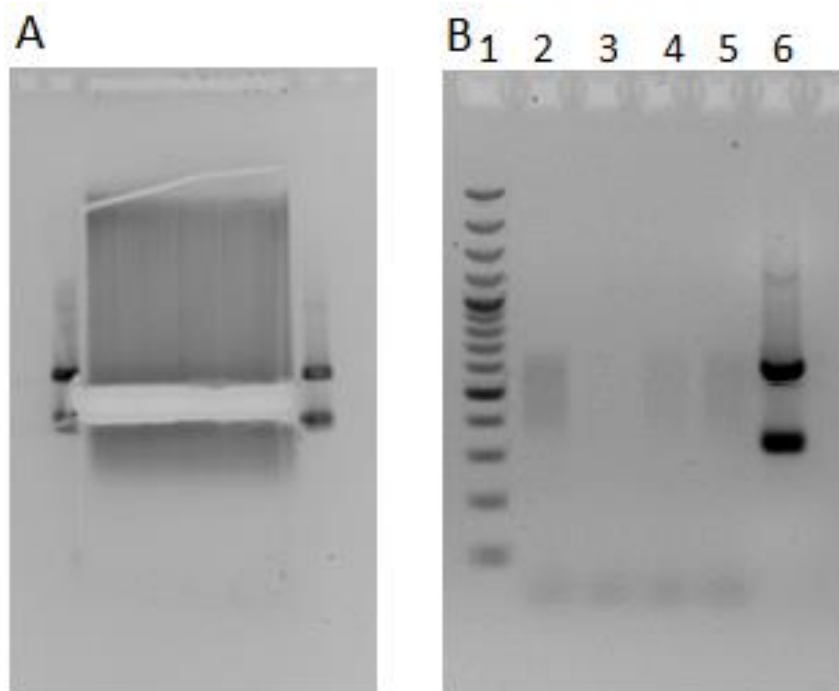


Figure 6.4: Gel images of the third (control) ddRAD library that was constructed from four fin samples in Nile tilapia. (A) Library size selection gel loaded with 320-590 bp markers to detect the size of interest, (B) represents the results of test PCRs where 100 bp GeneRuler, amplicon of 16X PCR with 0.5 μ L template, NTC, amplicon of 13X PCR with 2 μ L template, amplicon of 14X PCR with 1.5 μ L template and 320-590 bp markers that were previously used for the size selection of the library were loaded respectively from lane 1 to 6.

6.3.4 Troubleshooting

Once the control parental library proved that the variation observed in the sequencing outcome was not correlated with the usage of the multi-pipette (see Fig 6.3) and/or consistent failure of the restriction digestion which gave rise to either unusual size distribution towards bigger fragments or lower yield in the final library, the focus was then shifted to attempt to identify the problem by applying comprehensive troubleshooting procedures by testing: (i) adaptor & barcode sets, (ii) purification kit (PB buffer and 3Mm NaAc concentrations) and (iii) restriction enzymes specificity and genomic DNA.

6.3.4.1 *Adaptor & barcode test*

Figure 6.5A evidently demonstrates the results of adaptor set used in the previously failed ddRAD libraries in Nile tilapia was not due to adaptors or the ligation step, both stock and the aliquots (due to freeze-thaw steps) did not show any sign of detectable degradation. This was the first tested candidate for the reduced yield obtained in the libraries.

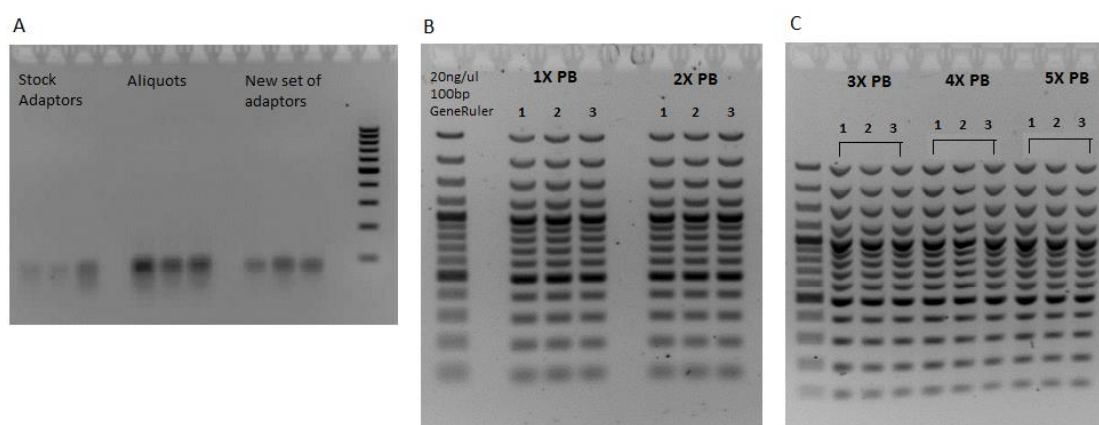


Figure 6.5: Troubleshooting steps carried out to identify the reason for reduced yield observed in ddRAD libraries constructed in Nile tilapia. (A) 2.5% agarose gel to assess any sign of degradation in adaptor sets, while images (B) and (C) shows the results of recovery of GeneRuler DNA from purification (Qiagen) kit with varying levels of PB buffer concentrations on 1.2% agarose gel. Lanes denotes as 1 had no NaAc, while lanes denoted as 2 had optimal NaAc concentration (1.5 μ L, colour change depending on pH was observed) and lanes denoted as 3 had a higher NaAc concentration (3.5 μ L).

6.3.4.2 *Purification kit test*

A brand-new Qiagen PCR purification kit was used for both first and the second ddRAD library construction in Nile tilapia as well as testing genomic DNA recovery with varying levels of PB buffer concentration alongside with the 3Mm NaAc. The manufacturer's recommendation is to use 5x more PB buffer for 1x gDNA to be purified with an optimum of 1.5-2.5 μ L 3Mm NaAc by observing colour change (bright yellow is recommended for maximum binding capacity to column based purification kit). In total

20 ng/ μ L 100 bp GeneRuler purified using varying levels of PB buffer from 1x to 5x concentrations with zero, optimal (1.5 μ L) or high (3.5 μ L) levels of 3 Mm NaAc processed by individual columns revealed that the recovery from the columns were almost 100% (Fig 6.5B-C; quantifications of the purified DNA, not shown here, revealed high correlation in terms of yield but no significant difference observed with varying concentrations of PB and/or 3Mm NaAc. This was not in accordance with the manufacturer's recommendations or concentration of NaAc which is a pH balancing reagent ensures high binding capacity to the column. Furthermore, increased PB concentrations did not help to get rid of smaller fragments as the manufacturer claims (note no difference in terms of intensity of smaller fragments - 100 and 200 bp - compared to same marker loaded to the first lane of gel B and C, Fig 6.5).

6.3.4.3 *Restriction enzymes specificity*

Once any degradation of reagents was ruled out as an explanation for reduced yield observed in ddRAD libraries constructed in Nile tilapia, attention was shifted towards the restriction digestion step. In the test digestion steps, the restriction reaction mix was identical to the ddRAD library procedure apart from increased genomic DNA (50 ng DNA per sample as opposed to 21ng) so that the digestion profiles could be better visualised. Figure 6.6 demonstrates the restriction digestion profiles from 2 fin clips and 2 larvae from which genomic DNA was freshly extracted. The restriction profile of *SbfI* enzyme with 8 bp recognition site, being a rare cutter, was a higher molecular weight distribution, while *SphI* enzyme with 6 bp recognition site, being the more common cutter, produced a broader smear on the gel (Fig 6.6). Significantly less digestion was observed in both larvae compared to fin clips in all reactions: *SbfI*, *SphI* and double digest

reaction mixes *SbfI*+*SphI*. The orange box represents the fragments of interest for ddRAD library procedures (note the lack of smear in larval samples compared to parental DNA extracted from fin clips). The efficiency of *SbfI*&*SphI* restriction digestion was also tested with different available master mixes (NEB4, CutSmart+1mM DDT; results not shown). CutSmart® Buffer performed the best as suggested by the manufacturer, hence the digestion of larvae and fin DNA was not reduced by the choice of master mix.

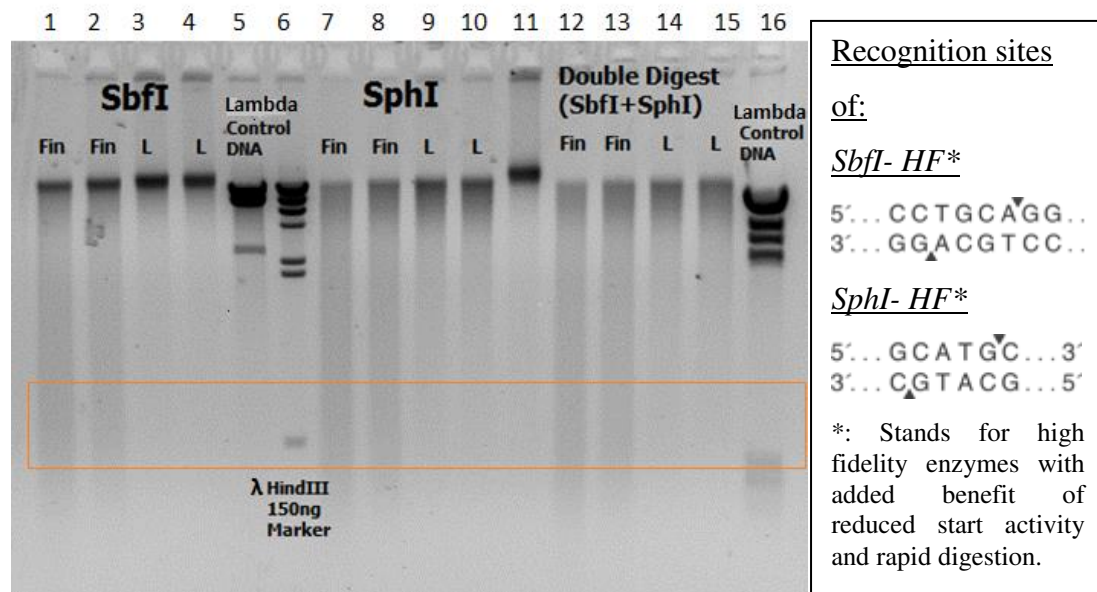


Figure 6.6: Restriction digestion profile of DNA from 2 fins and 2 larvae (in both groups gDNA were freshly extracted) on 1.1% agarose gel. The orange box represents the fragment range (300-700 bp) that is of interest for regular ddRAD library preparation. Higher concentration of (150ng) λ HindIII marker was loaded to lane 6 so as to observe 500 bp fragments to define region of interest. Lane 5 represents the Lambda DNA digested with *SbfI*, while lane 16 represents the Lambda DNA digested with *SbfI*+*SphI* as a control. Lane 11 presents the genomic DNA+MMix with no enzyme to observe any degradation in DNA level.

The next step was to try different restriction enzymes and observe their restriction pattern, to see whether such enzymes show any differences in terms of restriction pattern observed between parental (fin clips) and offspring (larvae) DNA. The enzyme finder tool of NEB (available at: <https://www.neb.com/tools-and-resources/interactive-tools/enzyme-finder>) was searched to find enzymes based on recognition sites similarity to *SbfI* and *SphI*. However, those with similar recognition sites were not available in the lab therefore

HindIII, *HaeIII* and *PstI*-HF enzymes were used for test digestion of fin and larvae genomic DNA. Figure 6.7 clearly shows that for *HaeIII* the digestion profile appeared similar (however this was not quantified) between genomic DNA extracted from fin or larvae while *HindIII* and *PstI* showed significant differences as observed for *SbfI* and *SphI* (See Fig 6.6) in the present study (see orange boxes highlighting the size of interest, 300 bp to 700 bp, or the bigger fragments on each lane for the ease of comparison of RE digested fragments).

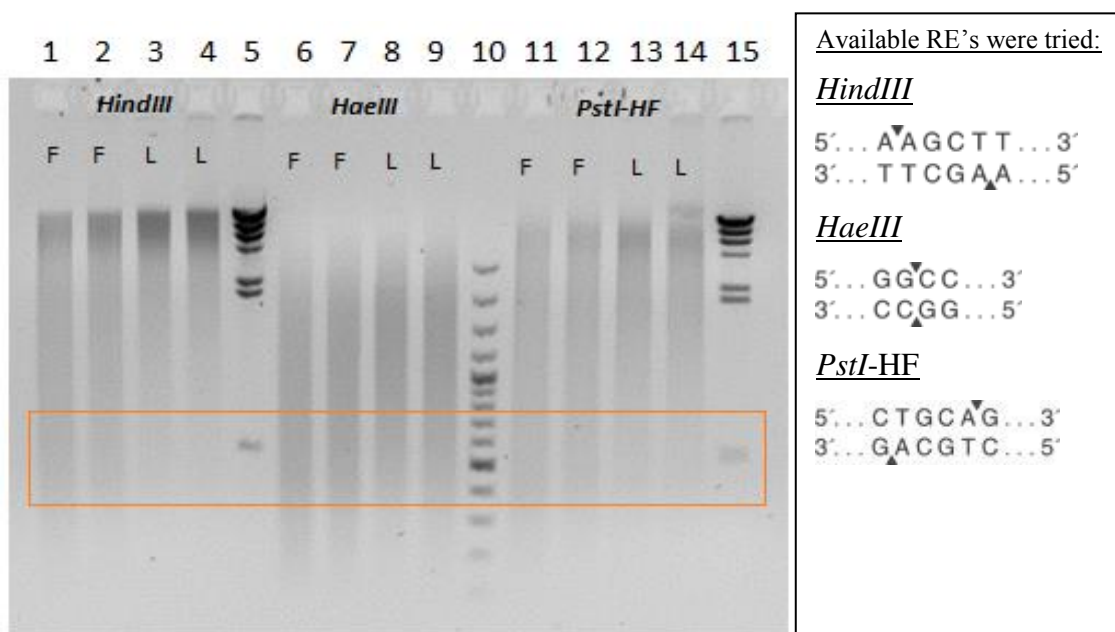


Figure 6.7: Restriction digestion profile of 2 fins and 2 larvae (in both groups gDNA were freshly extracted) run on 1.1% agarose gel using *HindIII*, *HaeIII* and *PstI*-HF. Lane 5 and 15 show a higher concentration (150ng) of λ *HindIII* marker so as to observe 500 bp size fragment while lane 10 shows 100 bp GeneRuler. The orange box represents the fragment range (300-700 bp) that is of interest for regular ddRAD library preparation.

6.3.4.4 Time series sampling

Taken together the restriction digestion profile differences observed with *SbfI*+*SphI* combination between genomic DNA extracted from fin clips, larva (Fig 6.6) and control library (see 6.3.3), confirm sufficient restriction digestion from four fin clips evident with 14x PCR cycles to enriched library, suggest that larval DNA has some sort of inhibition

against digestion by these enzyme combinations. One way to investigate this in depth was to produce another family and sample larvae in time-series as they grow so that both restriction digestion profile and optimised sampling time could be observed. To do that, a bi-parental family was produced (♀ 00-068D-OEDD x ♂ 00-068C-D6CA) and sampled up to 12 days AF. Table 6.3 shows the details of sampling which was based on the developmental stages of Nile tilapia described by Fujimura & Okada (2007).

Table 6.3: Schematic diagram of time series sampling regime carried out in a bi-parental Nile tilapia family.

Sampling	Period*	Stage*	Hour post fertilisation	Characteristic*
1st sampling	Pre-hatching	Stage 14	48hpf	Heart beat
2nd sampling	Hatching	Stage 17	96hpf	Jaw extension
3rd sampling	Post-hatching Early larva	Stage 22	8days AF	Swim bladder inflation
4th sampling	Late larva	Stage 25	10days AF	Yolk sac resorption Late larva
5th sampling	Early juvenile	Stage 28	12days AF	Free swimming activity

*: These parameters are based on Fujimura & Okada, (2007)

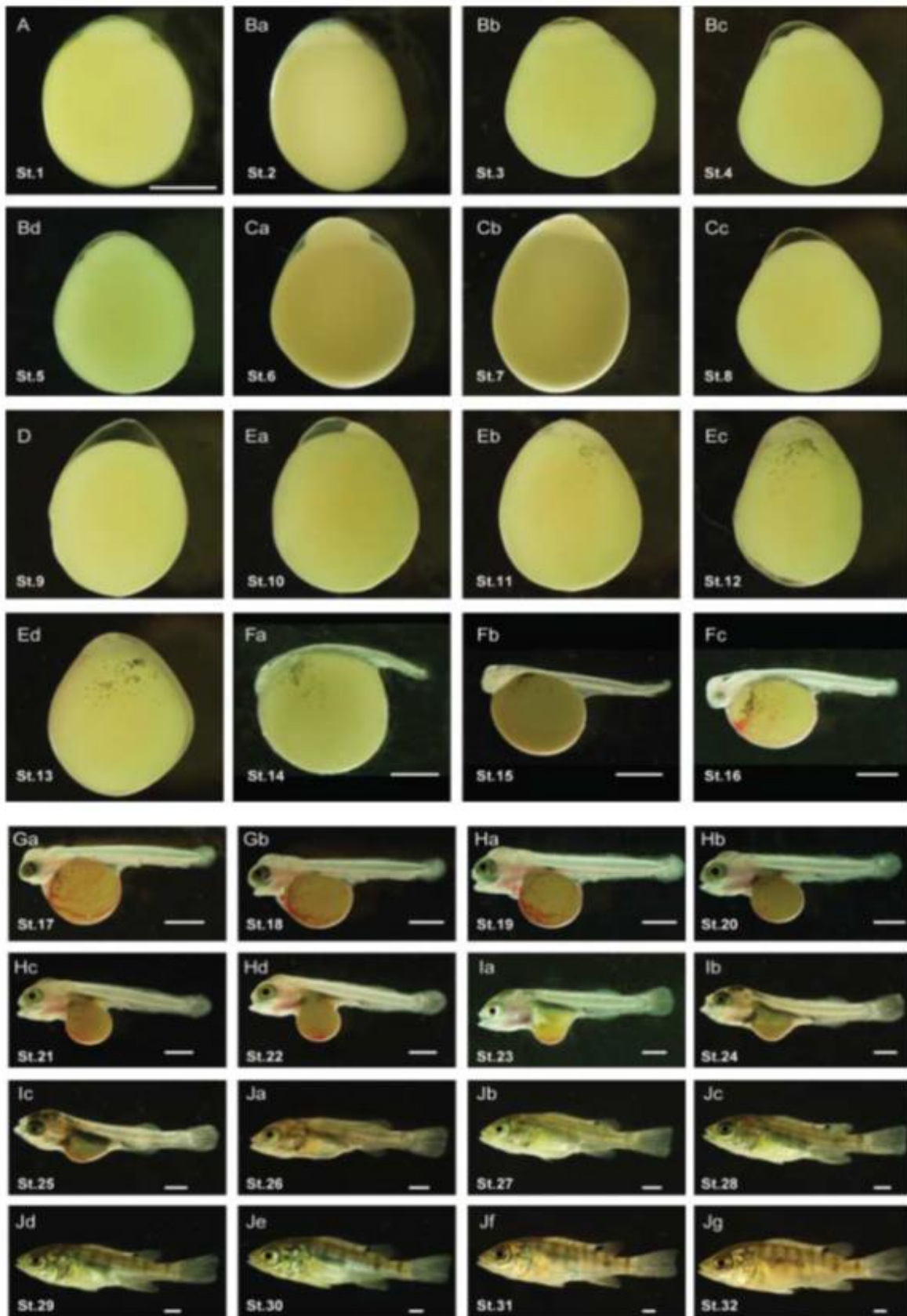


Figure 6.8: Image adapted from Fujimura & Okada (2007) to demonstrate time series sampling stages carried out in Nile tilapia (see Table 6.3 and the text for explanation).

Time series restriction digestion using *SbfI+SphI* revealed significant differences in resultant restriction fragments (Fig 6.9). Sufficient restriction digestion was only observed in genomic DNA that was extracted from fin clips (lane 2 to 5 in above gel, Fig 6.9). Larval DNA samples, on the other hand, failed to produce fragment that are of interest (300-700bp) for standard ddRAD library procedure so that adaptors could be ligated. However, the intensity of bigger fragments suggesting protection to restriction digestion or a potential global scale DNA methylation gradually decreased thus the restriction digestion profile improved as fish developed (note the difference in restriction digestion profile of pre-hatching to day-11, free swimming juvenile).

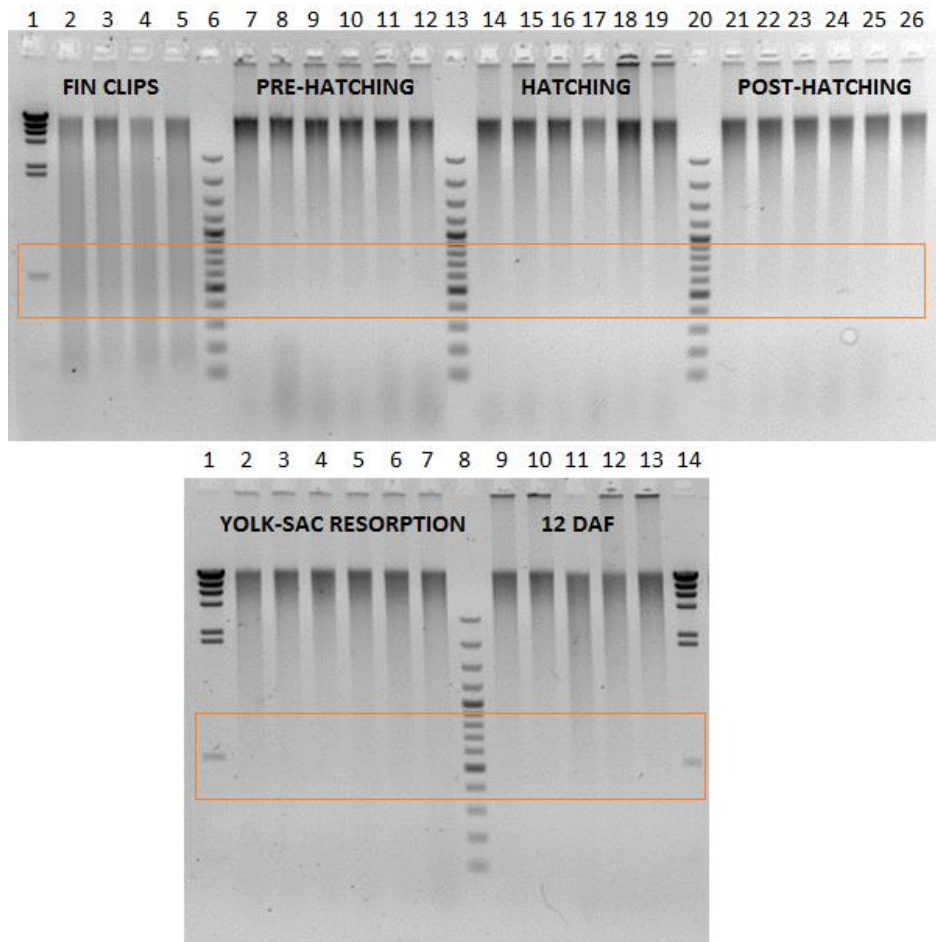


Figure 6.9: Double-digest restriction digestion profile of time series sampling carried out in Nile tilapia on 1.1% agarose gel. The orange box represents the fragment range of interest (300-700 bp) for regular ddRAD library procedure (These restriction enzyme digestion tests were carried out three times, one of which is shown here). Lane 1 of the upper gel and lane 1 and 14 of the below gels show higher concentration of *HindIII* (150 ng/ μ L) marker to observe 500 bp fragment size standard in both images while lanes 6, 13 and 20 of from the above gel and lane 8 from the below gel show 100 bp Gene Ruler.

6.4 Discussion

This study was primarily aimed to generate a genetic linkage map in an experimentally generated family of meiotic gynogenetic Nile tilapia (similar to Chapter 3 on seabass). A bi-parental control family from the same parental source was additionally added to the analysis to estimate the percentage of markers that segregate according to non-Mendelian inheritance. However, due to an early inhibition mechanism that was observed in Nile tilapia larval DNA, the restriction digestion was unsuccessful, regardless of the efforts made.

Although both enzymes (*SbfI* & *SphI*) used in ddRAD library procedure are not known to be methylation sensitive (for *dam*, *dcm* and *CpG* methylations) (<https://www.neb.com/products/epigenetics/methylation-sensitive-restriction-enzymes>), it can be argued that all restriction enzymes tested had CG bases adjacent in recognition sites which make them potentially likely to be affected by any methylation activity. DNA methylation is a common phenomenon where cellular processes including silencing of the specific parts of the genome are stimulated (see review by Goldberg et al., 2007). The most commonly seen and best characterised DNA methylation profile is CpG islands where some regions of genomes have higher densities of cytosine and guanine dinucleotides (Labbé et al., 2016). Furthermore the methylation profile is best studied in the early developmental stages, starting from gametogenesis (Primordial Germ Cells, PGCs) to early embryo development as the epigenome of the new organism is extensively reprogrammed (Hales et al., 2011). In an effort to identify fish epigenome profiles regarding to CpG islands, Han & Zhao (2008) performed a comparative study in five model fish species (tetraodon, stickleback, fugu, medaka and zebrafish) *in silico* and concluded that CpG islands greatly varied among the species, suggesting each species could have very divergent CpG numbers and densities. This could help to explain why

such inhibition digestion was only seen in Nile tilapia larvae given that we worked with Atlantic salmon and European seabass larvae at similar developmental stages with no inhibition of restriction enzyme digestion. However this would mean accepting DNA methylation as a causative of restriction digestion inhibition observed in the present study. Given that we neither have direct evidence to suggest DNA methylation nor has any other specific cause been identified within the timeframe of the present study, this would be only a speculation based on suggestive literature. One way to identify whether this inhibition observed in Nile tilapia larvae was associated with the DNA methylation would be to use DNA methylation sensitive restriction enzymes alongside non-sensitive restriction enzymes in a test panel of adult and larvae genomic DNA. This would either rule out or enhance the DNA methylation hypothesis. On the other hand using restriction enzymes that do not have CG in recognition site would also be informative. If such enzymes still produce the same pattern of inhibition in restriction digestion profile this would rule out DNA methylation as a likely cause. However as introduced above, DNA methylation studies have been focusing extensively on early developmental stages thus no records on adult methylome profiling are available in fish genomes as methylation is being associated with ageing effect, according to the author's limited knowledge on epigenomics in fish species (most of the knowledge still comes from mammalian studies in mice). Therefore it would be of interest to analyse the restriction digestion pattern of entire developmental stages in Nile tilapia including adult tissues with and without DNA methylated enzymes in future studies. This may highlight that it is more of a tilapia specific inhibition mechanism or potentially something else as opposed to being a universal DNA methylation.

Conflictingly, given the observations of even and complete restriction digestion profile in adult tissue, the result of the sequencing run was surprisingly biased towards larval DNA

where out of 6 (triplicated in both parents, see Fig 6.3A) parental DNA samples only two produced reads that were over the minimum threshold of 60,000 reads. However this ratio was higher in larval DNA samples (8 out of 9, 39 out of 59 meiotic gynogenetics produced reads that were over the minimum threshold of 60,000 reads). Although the reason for this variation is not clear, the fact that triplicated parental samples produced high variation in terms of reads derived from the same genomic DNA was a significant indicator of a bias in the library. Therefore no additional step was taken to re-sequence the same library to increase the depth of coverage.

An essential prerequisite of an effective ddRAD library is the sufficient fragmentation of genomic DNA of interest through the use of type II restriction enzymes. Insufficient fragmentation of the genome of interest, however, gives rise to random, bigger fragments which are not reproducible and generate unusual fragment size distribution, as seen in the first attempt of ddRAD library construction in Nile tilapia (section 6.3.1). In the case of failure at the restriction digestion step, adaptors carrying both barcodes (5 to 7 bp long molecular identifiers) and the sequencing primer site cannot be ligated. This subsequently reduces the fragments available as library template for the enrichment of the final library during bulk PCR procedure. Therefore longer PCR cycles can be required for sequencing. However longer PCR cycles (e.g. > 16) triggers bias in the final library. This probably contributes the variation observed in the sequencing of the second ddRAD library in Nile tilapia (section 6.3.2). Since early pooling was applied in the present study right after ligation of the adaptors as opposed to after individual PCRs per sample in the original protocol (Peterson et al., 2012), there was no way to equilibrate the number of reads produced at the end of the first sequencing run by simply going back to individual PCR amplicons for the potential subsequent sequencing run. However this modification was used to (i) accelerate the library wet-lab procedures from ten days to two days; (ii)

decrease pipetting thus reducing variation that would be introduced due to extra steps and to (iii) ensure PCR bias is kept at minimum by splitting reactions to as many tubes as possible (in half reaction volumes); and finally to (iv) save money and the consumables during purification steps with Qiagen and AmPure beads (e.g: in house-modified ddRAD protocol used in the present study processes a library using one column as opposed to using individual clean up per sample in the original protocols). PCR biases arises in NGS libraries are well documented in the literature (Rokas & Abbot, 2009; Pool et al., 2010; Davey et al., 2013; Arnold et al., 2013; Puritz et al., 2014).

Overall, detailed analysis of epigenetic processes in Nile tilapia larva was outside of the scope of the project and the expertise of researchers involved in the present study, as much as it was of scientific interest. Our initial aim for the current study could not be performed due to an unknown inhibition mechanism that was encountered in Nile tilapia larva which was identified by a series of systematic troubleshooting. It would be of great interest for future studies to identify the cause for such inhibitory mechanism encountered in Nile tilapia larvae. This study will be of help in terms of increasing awareness for researchers working in Nile tilapia genomics.

6.4.1 Conclusion

The present study attempted to generate a SNP-based genetic linkage map to locate centromere positions and identify more markers to distinguish between meiotic and mitotic gynogenetics, utilising the high-throughput power of NGS technologies. However as a result of an unknown early protection mechanism against restriction digestion faced in Nile tilapia larvae, genomic DNA of an experimentally produced family of meiotic gynogenetic larvae could not be efficiently restricted. These results were confirmed with a full-sib family produced and sampled in a time series later on. Regardless of the efforts

to overcome the challenge, the present study could not pin down the source of failure beyond it being related to DNA in larvae. Hence this study represents negative results which will however be useful for future studies in terms of increasing awareness while working with Nile tilapia larvae.

Chapter 7

General Discussion & Future Research Directions

General Discussion

The current study has attempted to gain new insights into the development of isogenic clonal fish lines in species of prime commercial interest of Europe, by using high-throughput sequencing technologies to address bottlenecks that have been encountered in this process. Within the scope of present study, species with (e.g: Atlantic salmon) and without (e.g: European seabass and Nile tilapia) genome duplication were studied. This chapter consists of an overall discussion on the main outcomes, strengths and limitations as well as future perspectives, organised at the species level.

7.1 European seabass (*Dicentrarchus labrax*)

Chapter 3 and 4 were dedicated to address the main bottlenecks observed in the development of isogenic clonal fish lines by (i) identifying markers at the distal end of the *D.labrax* chromosomes so as to identify informative markers to differentiate between mitotic and meiotic gynogenesis and by (ii) verifying genome-wide homozygosity in putative mitotic gynogenetics following an initial screening with a few microsatellite markers, respectively. In both experiments, as in all uniparental fish production protocol, the first focus was to confirm the lack of sire contribution in the putative gynogenetic progeny. This was achieved by screening segregation of paternal alleles. In the meiotic gynogenetics seabass experiment, 340 male informative markers were detected, located across all 24 linkage groups. None of these were detected in any of the 79 progeny, confirming the efficiency of the UV irradiation protocol (Peruzzi and Chatain, 2000) that was applied. A similar approach was applied to the analysis of putative mitotic gynogenetics (doubled haploids), where 4 families (F1, F3, F4 and F6) as well as two “orphans” (no parents available) were provided following an initial screening with 12 microsatellite loci. The absence of male alleles from paternal

informative SNP markers in each putative doubled haploid G1 family and in orphans (with uniformly homozygous genome) confirmed optimised UV irradiation (Peruzzi & Chatain, 2000) and the shock treatment protocol (Francescon et al., 2004) for *D.labrax*. Although protocol worked in limited number of fish (17 individuals out of 694 survival) a large number of fish were detected not to be doubled haploid by the microsatellite panel. To the best of author's knowledge this is the largest dataset has been used for the verification of isogenic clonal fish lines compared to (up to) tens (often less) of microsatellites routinely used up to now. Thus, successful and robust production of meiotic (Chapter 3) and mitotic gynogenetics (Chapter 4) was attained.

The high marker density achieved in both experiments not only helped to rule out sire contribution but also allowed construction of a SNP-based genetic linkage map based on the meiotic gynogenetic family. This map comprised a total of 764 SNPs spanning 1,252.02 cM with an average marker distance of 1.63 cM. Although the construction of a *de novo* genetic linkage map (as originally intended) was not possible, the reason for this is not entirely clear: this was either due to the nature of the data or the fact that genetic linkage mapping softwares are structured to take into account genetic contribution from both parents, as opposed to one phase of information available in the meiotic gynogenetic family. The seabass genome assembly was used to assign markers into linkage groups, but such a resource is not available for all species.

The main lesson learned was that parallel analysis of a bi-parental control group from the same family, with more informative meiosis so that the marker order attained could be combined with the heterozygosity values the meiotic gynogenetics, would have overcome this problem. Both datasets (meiotic gynogenetics and full-sib bi-parental controls) would be expected to contain the same set of markers. Previously, Guyomard et al. (2006) followed this approach to some extent (n = 60 in meiotic gynogenetics; n

= 60+60 in two F1 crosses between two isogenic lines) in rainbow trout. The strategy that Guyomard et al. (2006) used was based on manual ordering of limited amount of loci (not stated in the manuscript, pers.comm. R.Guyomard) in such a way that the number of recombination events was minimum, under the assumption of complete interference, in each linkage group in meiotic gynogenetic family. Thus the more likely interval in which centromeres lied was deduced from the gene-centromere distances (calculated as half the proportion of heterozygotes in meiotic gynogenetic progeny). Overall, there was not initial aim of constructing genetic linkage map based on meiotic gynogenetics data in Guyomard et al. (2006), despite such individual were used to help estimating centomeric regions along the chromosome arms of rainbow trout. This can be due to duplicated genome of salmonidae family members thus to eliminate complications.

In the present study a large number of SNP markers (n = 804 female heterogametic markers) were used initially to attempt generate a *de novo* genetic linkage map genome based on a meiotic gynogenetic family, therefore some problems could arise due to the change in the scale of big data. This has never been tried to the best of our knowledge. For example, a single false positive marker might imply in connecting LGs that are in fact different. This inflates some linkage groups with majority of markers are being falsely linked as encountered in the initial *de novo* genetic linkage map construction in meiotic gynogenetic family in seabass where LG 2 consisted of 483 markers out of 804 female heterogametic markers at LOD 14 which resulted in 25 LGs, closest to haploid chromosome number of *D.labrax*. This is because most genetic linkage mapping softwares, including OneMAP/R (Margarido, et al., 2007) used in the present study, use *transition* to link groups (if A is linked with B, and C is linked with B, automatically A and C are linked). The consequences of such cases following up with hundreds or

thousands of markers can have serious ordering problems and these cannot be controlled by simply increasing LOD value. One striking point needs to be highlighted at this point is that none of the markers in the biggest linkage group (LG 2) during initial *de novo* genetic linkage map construction shared anything in common neither heterozygosity value nor physical location close to one another, as one can initially think of. Regardless of mapping algorithms/interpretations used, this linkage could not be broken (Garcia, A.A.F. personal communication-OneMAP developer). Therefore, we used genome assembly grouping to construct genetic linkage in meiotic gynogenetic family and ordered markers within linkage groups.

The approach of assigning markers, first, into LGs based on reference genome assembly and then ordering markers using genetic linkage mapping software was later tested with the most current high density genetic linkage map in *D. labrax* (Palaiokostas et al., 2015b). This analysis revealed that our markers were a subset (15%) of those of Palaiokostas et al. (2015b) and the marker order corresponded across the 24 linkage groups, suggesting that this was a successful approach. Although 15% similarity might sound low, given the difference in library construction procedure (RADseq was used in the dense SNP map of Palaiokostas et al. (2015b) and ddRADseq was used in the present study) and different family origins (e.g. polymorphic loci in Palaiokostas et al. (2015b) dataset might not be polymorphic in present study) used, this was a reasonable proportion.

Additionally, high marker density achieved through the use of HTS technologies allowed identifying crossover locations along the chromosome arms of *D. labrax*; an average of 0.98 ± 0.12 , however multiple crossovers were also observed, suggesting interference is not complete. Marker-centromere recombination rates, ranged between zero and one providing an additional evidence on high coverage achieved in the present

study from proximal (centromeric) to distal (telomeric) regions along the genome of *D.labrax*. In total of 27 markers identified over 90% heterozygosity with 7 of them are being completely heterozygote in all 79 meiotic gynogenetic offspring, located on telomeric positions. These can be of primary interest for future studies once their diagnostic nature is verified. One point needs to be taken into account at this stage is that since such markers are derived from a family of meiotic gynogenetics they might not be necessarily shared among other families. Therefore a larger selection of markers might be needed for the initial validation of telomeric markers.

The evidence from meiotic seabass study suggests the existence of high crossover interference as reported by many aquatic species (Thorgaard, 1983; Danzmann & Gharbi, 2001; Morishima et al., 2001; Nomura et al., 2006; Martinez et al., 2008).

Large numbers of spontaneous meiotic gynogenetics were detected during the first phase of two-step verification study, by the initial microsatellite panel, and the results presented in the thesis highlighted the advantages of such a two-stage process and for a large number of DNA markers to distinguish between true meiotic and mitotic gynogenetics in the second stage. Until recently, most verification studies were based on a few loci. For example Francescon et al. (2005) used 5 microsatellite markers to detect the absence of paternal genetic contribution in experimentally induced progeny of meiotic gynogenetics in European seabass. Similarly, Ottera et al. (2011) genotyped a selection of 5 microsatellites so as to verify lack of paternal contribution in induced meiotic gynogenesis in Atlantic cod. However, as it was clearly demonstrated in Chapter 4, false positive(s), i.e. meiotic gynogenetics, can be detected among putative doubled haploid mitotic gynogenetic group. These false positives escaping from initial marker panel are triggered either by (i) limited number of markers used or the diagnostic power of microsatellites due to (ii) their position on the genome. In any

genotyping platform, regardless of the marker of choice, informative markers are the ones that are heterozygote in the parents. In the case of having monomorphic loci in parents will automatically decrease the diagnostic power of marker under investigation to zero. This in conjunction with the physical position of microsatellites located in low recombination centromeric regions might lead into false positive(s). Therefore the presence of false positives can only be detected by using large marker sets with almost genomically evenly spaces so that all regions (e.g: recombination low regions, centromeric or recombination high regions, telomeric) can be covered. For example, the only surviving progeny of F6 family was meiotic gynogenetic although it was initially being detected as mitotic gynogenetics based on 11 microsatellite loci. Detailed investigation of 11 microsatellite loci in F6 family revealed that there were only seven loci where female was heterozygous. Moreover these seven loci were located in the homozygosity blocks of linkage groups once their recombination was located on genetic linkage map generated from meiotic gynogenetic seabass family (Ch. 4; Table S2, available in electronic version).

European seabass is a prime importance aquatic species for Mediterranean. Therefore the need to develop genetic and genomic resources to reinforce future development of this species is well recognised. As a result of this, the first draft of genome assembly (Tine et al., 2014) publicly available alongside with well-established genetic linkage maps produced based on microsatellites, AFLPs and more recently SNP markers Chistiakov et al. (2005, 2008) and Palaiokostas et al. (2015b). The karyotype of European seabass consists of 24 subtelocentric-acrocentric chromosome pairs (Sola et al., 1993). A radiation hybrid map was generated to evaluate synteny analysis with model fish genomes and provided a complete gene map for the specie (Guyon et al., 2010) after the integration with the previous genetic linkage maps based on AFLP and

microsatellites by Chistiakov et al. (2005, 2008). However none of above genetic resources is able to provide diagnostic tool to differentiate between meiotic and mitotic gynogenetics due to non-localisation of centromeric regions, which was addressed in the present research. One remaining essential resource for the species is the establishment of isogenic clonal fish lines either through mitotic gynogenesis or androgenesis. Although induction of androgenesis could be more rapidly increased as some precocious males mature earlier as one year of age, the existence of mycosporin-like amino acids in the marine eggs acted as UV protection against varying doses of UV irradiation treatment thus lead into being ineffective at inactivating the maternal genome in European seabass (Colléter et al., 2014). Alternatively, there have been more promising efforts of producing effective mitotic gynogenetics (Bertotto et al., 2005) and meiotic gynogenetics as research-related studies (Peruzzi & Chatain, 2000). More recently, AquaExcel project (Horizon²⁰²⁰) (similar to forerunner project between 2011 and 2015 (FP7, EU) aims to establish isogenic clonal fish lines in species of commercial interest within Europe and Europeans seabass is one of the target species for the project in this regard.

7.2 Atlantic salmon (*Salmo salar*)

Although evidence regarding genome duplications in the course of evolution is widely accepted and well documented (Peer et al., 2009), the large portions of duplicated regions in genome post-WGD makes the process of identifying allelic forms of SNPs a difficult task (Hughes, 2007). The Atlantic salmon genome, as a representative of one of the most recent WGD events in vertebrates, almost three times larger than the average fish genome (3.4×10^9 bp), constitutes one of the most complex animal genomes (Danzmann et al., 2008). Although extant salmonids are in the process of

reverting back into stable diploid states through gene silencing or losing redundant copies of parts of the genome, still more than half of the genome is estimated to be in duplicated form (Lien et al., 2016).

Given these complications of the Atlantic salmon genome, a novel approach was applied to verify isogenicity in both G1 and G2 fish in Atlantic salmon propagated through mitotic and meiotic gynogenesis, by utilising the genome assembly to remove multi-copy loci following *de novo* analysis of short reads for building polymorphic loci was made. This constitutes one of the strengths of the study as opposed to assuming excess heterozygotes (e.g: >70%) indicating PSVs as used previously (Gonen et al., 2014; Houston et al., 2014). Detailed investigation of single-copy loci revealed 100% maternal allele transmission in G1 fish while varying levels of non-maternal alleles were detected persistently among the each putative clonal line, suggesting sub-optimal UV irradiation during propagation of second generation in salmon (unfortunately samples from the males that produced the sperm for irradiation were not kept). The results of ddRADseq analysis were in accordance with 27 (Norwegian + IOA, present study) microsatellite loci in G1 family confirming isogenic genome achieved in doubled haploid progeny. However, conflicting results were obtained in the next generation (G2) where ddRADseq identified varying levels of potential sire contribution persistent among G2 progenies. Although the 27 microsatellite loci covered 18 chromosomes out of 29 (with exceptions of 9 chromosomes; Ssal_04, Ssal_8, Ssal_10, Ssal_11, Ssal_18, Ssal_20, Ssal_21, Ssal_22, Ssal_23, Ssal_28 and Ssal_29), given the lack of gene-centromere map in Atlantic salmon, the diagnostic power of the microsatellites based on their location cannot be identified. Possible explanations are given as follows: (i) diagnostic power of microsatellites due to their position and/or common allele between parents or (ii) mutations occurring in primer binding site of microsatellites might have

occurred as homozygote (heterozygote to be detected as homozygote) yet there might be another reason(s) why these two marker technologies did not produce similar results in G2 families. However given relatively high number of microsatellite loci (27 markers) the latter scenario is less likely to occur in all microsatellite loci investigated. The fact that sire DNA was coming from a pool of two males (from which no DNA was provided) constituted difficulties during genotype calls. Therefore our strategy was to define any allele that did not match to maternal allele as a potential source for sire contribution. Given that no gene-centromere map is available in Atlantic salmon and marked differences in crossover frequencies make it difficult to locate the microsatellites so as to define their position relative to centromeres. To this end, the recent Atlantic salmon genetic linkage map by Gonen et al. (2014) proved an independent source for high coverage (throughout all chromosomes) achieved in the present study after a Blast search between two studies. An alternative way to test the genome-wide homozygosity would be the use of high density SNP chips (130K or 200K) available in salmon. Yet, the possibilities of having false positive homologs even with the stringent filtering process are considerable in SNP chip platforms. For example Dominik et al. (2010) used the 16.5K SNP chip designed for Atlantic salmon (Kent et al., 2009) and found high level of duplicated polymorphism (952 loci among 15,525 SNPs) for the Tasmanian Atlantic salmon populations.

Despite initial incompleteness of the draft genome in Atlantic salmon, the availability of a good quality genome assembly (Ssal_v4; AGKD00000000.4) made this study possible (by removing multi-copy loci). Recently reported repeat content of Atlantic salmon genome, 58-60 % (Lien et al., 2016) was observed in the present study, revealing high similarities with pike (*Esox lucius*) genome, phylogenetically the nearest sister non-duplicated relative group to Salmonids.

One of the limitations associated with Stacks, an open source pipeline for building loci from short sequence reads (Catchen et al., 2013), is the tendency of genotype calls being in favour of homozygotes. This can be best understood in an example. Assuming a locus that is heterozygous in both parents ($ab \times cd$) the genotype will be determined by the coverage achieved per each allele. If 170, 200, 9, 183 reads are available in parents, Stacks will be making a genotype call of $abxcc$ (by not accepting lower allele depth as being a valid alternate allele). Although such cases can be minimised through setting up appropriate parameters, given the nature of ddRAD library procedure such variations are inevitable cases on flow cell during bridge amplification in sequencing by synthesis. Regardless of the efforts made to start off with well quantified genomic DNA for each sample, PCR is likely to enrich some fragments more than the others. It is a well-known phenomenon that high GC content among loci is negatively correlated with read depth (Davey et al., 2013; Puritz et al., 2014). To this end, even removing PCR duplicates, which is the first candidate thought as introducing biases to the library, cannot solve this problem, hence requires pipelines that can deal with such complications due to the nature of workflow.

Future research work concerning verification of isogenicity in species with duplicated genome should consider the state and the availability of the reference genome assembly and involve genomic DNA of all the parents (preferably avoiding using pooled gametes). However false positive identifications (e.g: residual chromosome fragments not being detected) requires large numbers of genetic markers well spread along the genome of interest, thereby reducing the risk of false positives by ensuring accurate genotyping. Atlantic salmon is a species that has received tremendous amount of scientific, recreational and commercial interest. Thus genetic resources are well established and efforts are being currently made by many research groups to increase

genetic and genomic resources. One key resource can be facilitated is the establishment of isogenic clonal lines in species which was reviewed by Grimholt et al. (2009). In this regard, Atlantic salmon is one of the target species for AquaExcel project (Horizon²⁰²⁰) for the establishment of isogenic clonal lines in species of such value.

7.3 Nile tilapia (*Oreochromis niloticus*)

The idea in the Nile tilapia chapter was similar to that for meiotic European seabass: to generate a SNP-based genetic linkage map to locate centromere positions and identify more markers to distinguish between meiotic and mitotic gynogenetics by utilising high-throughput power of ddRADseq technique. However due to a restriction digestion inhibition in the early developmental stages of Nile tilapia, genomic DNA of larvae could not be digested sufficiently compared to homogenous and even digestion pattern observed in adult genomic DNA. Troubleshooting confirmed this reduced digestion profile in a full-sib family sampled throughout the development (from 48 hrs AF to 12 DAF), thus the initial objectives of the study could not be addressed.

Although the restriction enzymes used in the study were not known to be methylation sensitive, the fact that all enzymes had adjacent GC regions would make it possible that they could be affected by the most common form of epigenetic regulation (*CpG* islands) in mammals (Labbé et al., 2016). However given that there was no clear evidence for this from the study, it could only be speculated that DNA methylation was one of the possible causes of such inhibition in Nile tilapia larvae. One interesting aspect worth noticing was the lack of this inhibitory mechanism in both Atlantic salmon and European seabass larval DNA. Throughout the project we worked with early larvae (free swimming stage in European seabass (10 DAH) and up to 800 ddays larvae (yolk-sac larvae) in the Atlantic salmon. However no differences were observed in the

restriction digestion profile in both species and ddRAD libraries were successfully constructed using the same protocol.

One way to investigate this further would be to use a similar panel of samples including both adult and larval genomic DNA in different developmental stages and observing restriction digestion profile with DNA methylation sensitive enzymes. As much as it presents an interesting field of research, the results of this study were unexpected and were outside of the scope of the present PhD thesis. A series of troubleshooting steps was performed but did not lead to any conclusion beyond an unknown inhibitory restriction digestion mechanism that was clearly observed in Nile tilapia larva compared to adult genomic DNA and is of interest for researchers working in Nile tilapia genomics.

7.4 The role of HTS technologies on the future of Isogenic clonal fish line development

Development of isogenic clonal lines is a complex and expensive exercise. Fish species with their diverse reproductive behavior, variations observed in gamete quality and environmental factors affecting husbandry practises affect the quality of the resultant progeny. Close monitoring is required in each step of the development of isogenic clonal fish lines so as to avoid inclusion of biparental and/or meiotic gynogenetic fish into putative isogenic stocks. Up to now, many research groups have been using limited numbers of markers (less than 10 in most species), some of which have verified for their diagnostic power but some not. ddRADseq was proved to be successful for the process of development of isogenic clonal fish lines in both European seabass and Atlantic salmon (with a duplicated genome). Taken all together, it is recommended to apply this scale of marker genotyping for future screening of isogenic clonal fish lines.

Despite the advantages of isogenic clonal fish lines mainly for aquaculture related research, the maintenance of such lines is difficult due to issues related to husbandry, cost and inbreeding depression. Additionally, most commercial fish species have a generation interval of 2-3 years and in some species up to 4 years (i.e: Atlantic salmon and Atlantic halibut). Efforts to accelerate the generation of isogenic European seabass by applying androgenesis encountered difficulties apparently due to pelagic marine eggs having UV protective compounds, creating problems to inactivate the egg nuclear genome (Colléter et al., 2014). Taken all together, development of isogenic clonal lines requires dedicated research and well-equipped facilities so as to provide good husbandry for DH fish. To this end, INRA experimental farm (PEIMA, Sizun, France) in France under E. Quillet's maintenance responsibility, and Aquaculture Core Facility, centre for reproductive biology (Washington State University) in USA under S. Ristrow's maintenance responsibility constitute encouraging examples in isogenic fish lines in rainbow trout. However, there is a clear need for longer term funding for such lines to be developed in other species and maintained under close monitoring. The AquaExcel project (Horizon²⁰²⁰) and its forerunner (EU, FP7 between 2011-2015) aim to establish isogenic clonal fish lines for research-related use in species of commercial interest for Europe.

7.5 The role of HTS technologies on the future developments of aquaculture genomics

The present study relied on using hundreds to thousands of molecular markers, in the form of SNPs, for investigation and verification of isogenic clonal fish line production techniques. These could not have been analysed based on phenotypic variation,

therefore reliable genotypic data is needed. However genotypic variation of a given species is not observed in every genomic location but only in those landmarks termed as genetic markers (Liu & Cordes, 2004; Liu 2011).

One of the main limitations of aquaculture genetics has been the availability of limited number of markers and the cost related to identify and genotype such markers from various species. As the number of genetic markers has increased, the trait(s) under investigation can be analysed better as denser genetic marker information covers the whole genome of interest. This is where HTS technologies come in, with large numbers of genetic markers.

The potential arises from supreme scale change in genetic marker density not only enables comprehensive genome-wide studies (Davey et al., 2011) but also empowers Marker Assisted Selection (MAS) to be applied in aquatic species with increased accuracy (Sonesson, 2011). MAS, different than conventional selection, involves the addition of genotypic information in the form of genetic markers to the phenotypic data so as to better monitor and increase the selection response (Yue, 2013). It is a very good tool for traits such as disease resistance or fillet quality, where selection response cannot be directly measured on selection candidates but is measured in siblings of candidates. Currently with the advances in HTS technologies SNPs covering the entire genome are readily available in realistic costs and timeframes. Thus making identification of markers that are linked to a trait of interest feasible for many aquatic species. This said, aquaculture still presents challenges on implication of sophisticated breeding programmes due to the difficulty of maintaining stocks: for example single-pair matings are still relatively uncommon (Gjerdem & Robinson, 2014). However the added benefits of molecular markers can be facilitated in the form of shortened generation interval (Sonesson, 2011).

Applications of MAS with direct impact on the industry can be examined in two recent cases: (i) lymphocystis disease–resistance in Japanese flounder and (ii) IPN resistance in Atlantic salmon. Fuji et al. (2007) identified a microsatellite locus, *Poli9-8TUF*, associated with disease resistance in the Japanese flounder and produced a new population of progeny (by crossing between a female homozygous for favourable, disease resistance allele and a male selected for higher growth and body shape without a resistance allele so that all progeny would have resistant-heterozygote genotype for disease and perform better under commercial operations) by applying MAS with *Poli9-8TUF* locus. Challenge test confirmed the resistance of MAS applied progeny in response to lymphocystis disease compared to control family. Ozaki et al. (2012) evaluated the effect of MAS market penetration rate of resistant Japanese flounder as 35% in Japan in 2012. Houston et al. (2008, 2012) and Moen et al. (2008) identified a major QTL responsible for resistance to IPN in Atlantic salmon. Commercialised resistant salmon eggs through MAS (QTL-innOva®) performed better in the field under commercial operations. These two studies constitute the first applications of MAS in aquaculture breeding with major impact on commercial operations while more applications of MAS is well established in terrestrial animals (i.e: *MC4R* gene Houston et al., 2004). Although neither mapping nor applications of MAS are as advanced as in terrestrial animals compared to aquatic species, given the potential that NGS offers with thousands of markers that are almost evenly distributed along the genome of interest in realistic costs, the upcoming years are expected to be fruitful for direct application in the aquaculture industry (Martinez, 2007).

Rapid advances achieved in sequencing technology have led to rapid advances in several aquatic species where there was little or no genetic background information available previously. Several fish genomes have been sequenced, including both species

of economic (e.g: Atlantic salmon, common carp, European seabass, Nile tilapia) and evolutionary interest (lamprey as a out-group to ancestral vertebrate genome; shark, as a model for cartilaginous fish) (Lien et al., 2016; Xu et al., 2014; Tine et al., 2014; Brawand et al., 2014; Smith et al., 2013; Venkatesh et al., 2014).

The availability of fully sequenced fish genomes brings new opportunities for primarily aquaculture-related research and for the aquaculture industry to develop successful applications. Genome assemblies will be of use for developing new vaccines by targeting specific regions of the genomes (Locke et al., 2008), improve feeding by understanding gene regulations and physiology related to alternative diets (Glencross et al., 2015) and more efficient and targeted selective breeding through MAS with increased responses (Ozaki et al., 2012).

As demonstrated throughout the present research, ddRADseq offers excellent, fast and economic solution for any study requiring medium-scale genotyping. The practical aspect of ddRADseq lies in the flexibility of increasing sample size by simply designing specific combinations of adaptors that can accommodate larger numbers of individuals. This, in conjunction with different combinations of restriction enzymes would be most interesting direction for future research to be undertaken. It is likely that ddRADseq and other variations on RADseq will continue to be methods of choice in the forthcoming years, however other platforms available also offer more flexibility to researcher depending on the objectives of studies.

7.5 General summary

In summary, the main outcomes of the current study were as follows:

- Verified optimised sperm UV inactivation protocol for European seabass in

genome-wide scale, for the first time, confirming uniparental inheritance in both meiotic (*Chapter-3*) and mitotic (*Chapter 4*) gynogenetic experimental groups.

- Construction of, for the first time, a genetic linkage map based on a meiotic gynogenetic family in European seabass with a marker-centromere map, thus identifying centromere positions and large numbers of telomeric markers with high recombination frequencies (*Chapter-3*) as a step for development of isogenic clonal fish lines to detect and eliminate untargted occurrence of meiotic gynogenetics in mitotic gynogenetic groups.
- Verified optimised sperm UV inactivation protocol for Atlantic salmon in genome-wide scale, for the first time, confirming uniparental inheritance in clone founders (G1) while detecting varying levels of sire contribution in the next generation (G2) (*Chapter-5*).
- Detected an unknown inhibitory protection mechanism against restriction digestion in the early developmental stage of Nile tilapia larvae (*Chapter 6*).
- Validated the efficacy of HTS technologies as an improved way of detecting potential residual contribution from irradiated gametes in both fish species with and without duplicated genomes.
- Provided the first evidence concerning successful application of HTS technologies in species with ancestral tetraploid origin in isogenic clonal fish lines development.

Overall, the HTS platform used throughout the analysis of the present thesis proved the utility of such platforms to meet the objectives of the present research. However a good quality reference genome assembly was essential while working with a species of ancestral tetraploidy (to identify and eliminate duplicated loci), and as an aid in developing a linkage map based on a meiotic gynogenetic family (for initial grouping of

markers). Although these platforms are a cost-effective way of genotyping, they are not practical where very large numbers of fish need to be genotyped for genome-wide homozygosity. To this end, two-step selection method applied in the verification of putative doubled haploid mitotic gynogenetics in European seabass provided a realistic pilot study (Chapter 4). The initial mass selection of the mitotic gynogenetics was carried out with previously validated microsatellite panels with high recombination frequencies. This resulted significantly reducing the number of fish to be screened in the second round of verification by using more genomically comprehensive marker technologies such as ddRADseq. By doing so, a total of 694 putative mitotic gynogenetics were scaled down to 30 mitotic gynogenetics based on initial selection of 12 microsatellite loci, by decreasing both the timeframe and the cost of verification. To this end, future research concerning verification of isogenic clonal fish lines can be accelerated by facilitating two-step verification approach where markers are available for initial genotyping followed up with more accurate and fast-forward screening of genome-wide scale. This would not only detect possible residual chromosome fragments from irradiated gametes but also detect any false positives that might arise in the doubled haploid group with less genomically comprehensive marker technologies.

References

- Ahituv, N., Rubin, E. M. & Nobrega, M. A. (2004). Exploiting human - Fish genome comparisons for deciphering gene regulation. *Human Molecular Genetics*, (13): 261–266.
- Aljanabi, S.M. & Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, 25(22), 4692–4693.
- Allendorf, F.W. (1978). Protein polymorphism and the rate of loss of duplicate gene expression, *Nature* 272: 76 – 78.
- Allendorf, F.W., Bassham, S., Cresko, W. A, Limborg, M. T., Seeb, L. W. & Seeb, J. E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *The Journal of Heredity*, 106(3), 217–27.
- Alsaqufi, A. S., Gomelsky, B., Schneider, K. J. & Pomper, K. W. (2012). Verification of mitotic gynogenesis in ornamental (Koi) carp (*Cyprinus carpio* L.) using microsatellite DNA markers. *Aquaculture Research*, 45(3), 410–416.
- Amores, A., Catchen, J., Ferrara, A., Fontenot, Q. & Postlethwait, J. H. (2011). Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, 188(4), 799–808.
- Anderson, J.L., Rodri, A.R, Braasch, I., Amores, A., Hohenlohe, P., Batzel, P. & Postlethwait J.H. (2012). Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. *PloS One*, 7(7), e40701.
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Arai, K., Onozato, H., Yamazaki, F. (1979). Artificial androgenesis induced with gamma irradiation in masu salmon, *Oncorhynchus masou*. *Bulletin of the Faculty of Fisheries*, Hokkaido University 30, 181–186.
- Arai, K., Masaoka, T. & Suzuki R. (1992) – Optimum conditions of UV irradiation for genetic inactivation of loach eggs – *Nippon Suisan Gakk.* 58: 1197-201.
- Arai, K. (2001). Genetic improvement of aquaculture finfish species by chromosome manipulation techniques in Japan. *Aquaculture*, 197(1–4), 205–228.
- Arnold, B., Corbett-Detig, R. B., Hartl, D. & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, 22(11), 3179–90.
- Babbucci, M., Ferraresso, S., Pauletto, M., Franch, R., Papetti, C., Patarnello, T., Carnier, P. and Bargelloni, L. (2016). An integrated genomic approach for the study of mandibular prognathism in the European seabass (*Dicentrarchus labrax*). *Nature Scientific reports*, 6: 38673.
- Babiak, I., Dobosz, S., Goryczko, K., Kuzminski, H., Brzuzan, P. & Ciesielski, S. (2002). Androgenesis in rainbow trout using cryopreserved spermatozoa: the effect of processing and biological factors. *Theriogenology*, 57(4), 1229–49.

- Baird, N., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E.U., Cresco, W.A. & Johnson, E. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, 3(10), e3376.
- Balding, D., Bishop, M. & Cannings, C. (2007). *Handbook of Statistical Genetics* (Vol. 1, p. 1533). John Wiley & Sons Ltd.
- Bayne, C.J., Gerwick, L., Wheeler, P.A. & Thorgaard, G.H. (2006). Transcriptome profiles of livers and kidneys from three rainbow trout (*Oncorhynchus mykiss*) clonal lines distinguish stocks from three allopatric populations. *Comparative Biochemistry and Physiology. Part D, Genomics & Proteomics*, 1(4), 396–403.
- Beck, J.A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J. T., Festing, M.F. & Fisher, E.M. (2000). Genealogies of mouse inbred strains. *Nature Genetics*, 24(1), 23–25.
- Ben-Dom, N., Cherfas, N.B., Gomelsky, B., Avtalion, R.R., Moav, B. & Hulata, B. (2001). Production of heterozygous and homozygous clones of common carp (*Cyprinus carpio* L.): Evidence from DNA fingerprinting and mixed leukocyte reaction. *Israeli Journal of Aquaculture – Bamidgeh*. 53:89–100.
- Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Albetti, A., Aury, J.M., Louis, A., Dehais, P., Bardou, P., Monfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G.H., Boussaha, M., Quillet, E., Guyomard, R., Galiana, D., Bobe, J., Volff, J.N., Genet, C., Wincker, P., Jaillon, O., Roest, C., Hugues, R.C. & Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, 5, 3657.
- Bertioli, D.J., Ozias-Akins, P., Chu, Y., Dantas, K. M., Santos, S.P., Gouvea, E.G., Guimaraes, P.M., Leal-Bertioli, S.C.M., Knapp, S.J. & Moretzsohn, M.C. (2014). The Use of SNP Markers for Linkage Mapping in Diploid and Tetraploid Peanut. *G3* (Bethesda, Md.). 4(1): 89-96.
- Bertotto, D., Cepollaro, F., Libertini, A., Barbaro, A., Francescon, A., Belvedere, P., Barbaro, J. & Colombo, L. (2005). Production of clonal founders in the European sea bass, *Dicentrarchus labrax* L., by mitotic gynogenesis. *Aquaculture*, 246(1-4), 115–124.
- Blázquez, M., Zanuy, S., Carillo, M. & Piferrer, F. (1998). Effects of Rearing Temperature on Sex Differentiation in the European Sea Bass (*Dicentrarchus labrax* L.). *Journal of Experimental Zoology*, 216; 207–216.
- Bongers, A.B.J., In't Veld E.P.C., Avo-Hashema, K., Bremmer, I.M., Eding, E.H., Komen, J. (1994). Androgenesis in common carp (*Cyprinus carpio* L.) using UV irradiation in a synthetic ovarian fluid and heat shocks. *Aquaculture* 122: 119–132.
- Bongers, A.B.J., Bovenhuis, H., Van Stokkom, A. C., Wiegertjes, G. F., Zandieh-Doulabi, B., Komen, J. & Richter, C.J.J. (1997a). Distribution of genetic variance in gynogenetic or androgenetic families. *Aquaculture*, 153(3-4), 225–238.
- Bongers, A.B.J., Benayed, M.Z., Doulabi, B.Z., Komen, J. & Richter, C.J.J. (1997b) Origin of variation in isogenic, gynogenetic and androgenetic strains of common carp, *Cyprinus carpio*. *Journal of Experimental Zoology*, (277) 72–79.
- Bongers, A.B.J., Zandieh-Doulabi, B., Voorthuis, P. K., Bovenhuis, H., Komen, J. & Richter, C.J.J. (1997). Genetic analysis of testis development in all-male F1 hybrid strains of common carp, *Cyprinus carpio*. *Aquaculture*, 158, 33–41.

- Bongers, A.B.J., Sukkel, M., Gort, G., Komen, J. & Richter, C.J.J. (1998). Development and use of genetically uniform strains of common carp in experimental animal research. *Lab Animals*, 32(4), 349–363.
- Bongers, A.B.J., Richter, C.J.J. & Komen, J. (1999). Viable Androgenetic YY Genotypes of Common Carp (*Cyprinus carpio* L.). *Journal of Heredity*, 90(1), 195–198.
- Bourret, V., Kent, M. P., Primmer, C. R., Vasemägi, A., Karlsson, S., Hindar, K., McGinnity, P., Verspoor, E., Bernatches, L. & Lien, S. (2013). SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, 22(3), 532–51.
- Boutin-Ganache, I., Raposo, M., Raymond, M., Deschepper, C. F. (2001). M13-tailed primers improve the readability and usability of microsatellite analyses performed with two different allele-sizing methods. *BioTechniques* 31: 25–28.
- Braasch, I. & Postlethwait, J. H. (2012). Polyploidy and Genome Evolution. (P.S. Soltis & D. E. Soltis, Eds.). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Brawand, D., Wagner, C. E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H.J., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., Baroiller, J.F., Barloy-Hubler, F., Berlin, A., Bloomquist, R., Carleton, K.L., Conte, M.A., D'Cotta, H., Eshel, O., Gaffney, L., Galibert, F., Gante, H.F., Gnerre, S., Greuter, L., Guyon, R., Haddad, N.S., Haerty, W., Harris, R.M., Hofmann, H.A., Hourlier, T., Hulata, G., Jaffe, D.B., Lara, M., Lee, A.P., MacCallum, I., Mwaiko, S., Nikaido, M., Nishihara, H., Ozouf-Costaz, C., Penman, D.J., Przybylski, D., Rakotomanga, M., Renn, S.C.P., Ribeiro, F.J., Ron, M., Salzburger, W., Sanchez-Pulido, L., Santos, M.E., Searle, S., Sharpe, T., Swofford, R., Tan, F.J., Williams, L., Young, S., Yin, S., Okada, N., Kocher, T.D., Miska, E.A., Lander, E.S., Venkatesh, B., Fernald, R.D., Meyer, A., Ponting, C.P., Streelman, J.T., Lindblad-Toh, K., Seehausen, O. & Di Palma, F. 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, vol 513, 7518, pp. 375-381.
- Brown, J. K., Taggart, J. B., Bekaert, M., Wehner, S., Palaiokostas, C., Setiawan, A. N., Symond, J.E. & Penman, D. J. (2016). Mapping the sex determination locus in the hāpuku (*Polyprion oxygeneios*) using ddRAD sequencing. *BMC Genomics*, 17(1), 448.
- Brown, K. H., Lee, R. W. & Thorgaard, G. H. (2006). Use of androgenesis for estimating maternal and mitochondrial genome effects on development and oxygen consumption in rainbow trout, *Oncorhynchus mykiss*. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, 143(4), 415–21.
- Brown, K.H. & Thorgaard, G.H. (2002). Mitochondrial and nuclear inheritance in an androgenetic line of rainbow trout, *Oncorhynchus mykiss*. *Aquaculture*. 204: 323–335.
- Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C. & DePristo, M.A. (2012) Pacific bioscience sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, 13, 375.
- Cartwright, D. A, Troggio, M., Velasco, R. & Gutin, A. (2007). Genetic mapping in the presence of genotyping errors. *Genetics*, 176(4), 2521–7.
- Castillo-Davis, C.I., Hartl, D.L. & Achaz, G. (2004). cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res.* 14, 1530–1536.

- Castro, J., Bouza, C., Sánchez, L., Cal, R.M., Piferrer, F., Martínez, P. (2003). Gynogenesis assessment by using microsatellite genetic markers in turbot (*Scophthalmus maximus*). *Mar. Biotechnol.* 5, 584–592.
- Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W & Postlethwait, J. H. (2011). Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Genes, Genomes, Genetics)*, 1(3), 171–82.
- Catchen, J., Bassham, S., Wilson, T., Currey, M., O’Brien, C., Yeates, Q. & Cresko, W. A. (2013). The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, 22(11), 2864–83.
- Chapuis, M.P. & Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, 24(3), 621–31.
- Chavanne, H., Janssen, K., Hofherr, J., Contini, F., Haffray, P., Komen, H., Nielsen, E.E & Bargelloni, L. (2016). A comprehensive survey on selective breeding programs and seed market in the European aquaculture fish industry. *Aquaculture International*, 24(5), 1287–1307.
- Chistiakov, D. A., Hellemans, B., Haley, C. S., Law, A. S., Tsigenopoulos, C. S., Kotoulas, G., Bertotto, D., Libertini, A. & Volckaert, F. A. M. (2005). A microsatellite linkage map of the European sea bass *Dicentrarchus labrax* L. *Genetics*, 170(4), 1821–6.
- Chistiakov, D., Tsigenopoulos, C. S., Lagnel, J., Guo, Y. M., Hellemans, B., Haley, C. S., Volckaert, F. A. M. & Kotoulas, G. (2008). A combined AFLP and microsatellite linkage map and pilot comparative genomic analysis of European sea bass *Dicentrarchus labrax* L. *Animal Genetics*, 39(6), 623–34.
- Chourrout, D. & Quillet, E. (1982). Induced gynogenesis in the rainbow trout : sex and survival of progenies, production of all-triploid populations. *Theoretical and Applied Genetics*, 63, 201-205.
- Chourrout, D., Chevassus, B., Krieg, F., Happe, A., Burger, G. & Renard, P. (1986). Production of second generation triploid and tetraploid rainbow trout by mating tetraploid males and diploid females – potential of tetraploid fish. *Theories of Applied Genetics*. 72:193– 206.
- Chutimanitsakun, Y., Nipper, R. W., Cuesta-Marcos, A., Cistué, L., Corey, A., Filichkina, T., Johnson, E.A. & Hayes, P. M. (2011). Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics*, 12; 4.
- Colléter, J. (2015). Towards the development of clonal lines in the Europeans seabass (*D. labrax* L.): application of uniparental reproduction techniques with insights into seabass eggs. PhD thesis, University of Montpellier 2, available online.
- Colléter, J., Penman, D. J., Lallement, S., Fauvel, C., Hanebrekke, T., Osvik, R. D., Eilertsen, H.C., D’Cotta, H., Chain, B. & Peruzzi, S. (2014). Genetic inactivation of European sea bass (*Dicentrarchus labrax* L.) eggs using UV-irradiation: observations and perspectives. *PloS One*, 9(10), e109572.
- Crête-Lafrenière, A., Weir, L. K. & Bernatchez, L. (2012). Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. *PloS One*, 7(10), e46662.
- Crisp, D. T. (ed.) 2000. Trout and salmon: ecology, conservation and rehabilitation. Oxford Fishing News Books, UK.

- Danzmann, R. G. & Gharbi, K. (2001). Gene mapping in fishes: a means to an end. *Genetica* 111, 3–23.
- Danzmann, R.G., Cairney, M., Davidson, W.S., Ferguson, M.M., Gharbi, K., Guyomard, R., Holm, L.E., Leder, E., Okamoto, N., Ozaki, A., Rexroad, C.E., Sakamoto, T., Taggart, J.B. & Woram, R.A. (2005). A comparative analysis of the rainbow trout genome with two other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily : Salmoninae). *Genome*, vol 48, no. 6, pp. 1037.
- Danzmann, R.G., Davidson, E.A., Ferguson, M.M., Gharbi, K., Koop, B.F., Hoyheim, B., Lien, S., Lubieniecki, K.P., Moghadam, H.K., Park, J., Phillips, R.B., Davidson, W.S. (2008). Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (rainbow trout and Atlantic salmon). *BMC Genomics*, 9:557.
- Dave, G., (1993). Replicability, repeatability and reproducibility of embryo-larval toxicity-tests with fish. In: Soares, A.M.V.M., Calow. P. (Eds.), *Progress in Standardization of Aquatic Toxicity Tests with Fish*. Lewis Publishers, London.
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M. & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7), 499–510.
- Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, 22(11), 3151–64.
- David, L., Blum, S., Feldman, M.W., Lavi, U. & Hillel, J. (2003). Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. *Mol. Biol. Evol.* 20, 1425–1434.
- Davidson, W. S., Koop, B. F., Jones, S. J. M., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S. & Omholt, S. W. (2010). Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biology*, 11(9), 403.
- DeAngelis, M.M., Wang, D.G. & Hawkins, T.L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research*, Vol. 23, No. 22, 4742-4743
- Devlin, R. H. & Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture*, 208, (3-4): 191–364.
- Diter, A., Quillet, E., Chourrout, D., 1993. Suppression of first egg mitosis induced by heat shocks in the rainbow trout. *Journal of Fish Biology* 42, 777–786.
- Dominik, S., Henshall, J. M., Kube, P. D., King, H., Lien, S., Kent, M. P. & Elliott, N. G. (2010). Evaluation of an Atlantic salmon SNP chip as a genomic tool for the application in a Tasmanian Atlantic salmon (*Salmo salar*) breeding population. *Aquaculture*, 308, S56–S61.
- Dunham, R. A. (2004). *Aquaculture and Fisheries Biotechnology* (p. 367). Oxfordshire, UK: Cabi Publishing.
- Dunham, R.A. & Brummett, R.E. (1999). Response of two generations of selection to increased body weight in channel catfish, *Ictalurus punctatus*, compared to hybridization with blue catfish, *I. furcatus*, males. *Journal of Applied Aquaculture*, 9: 37–45.

- Durbeej, B. & Eriksson, L.A. (2002). Reaction mechanism of thymine dimer formation in DNA induced by UV light, *Journal of Photochemistry and Photobiology A:Chemistry*, 2002, 152, 95.
- Elmer, K. R. & Meyer, A. (2011). Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, 26(6), 298–306.
- Everett, M.V., Miller, M. R. & Seeb, J. E. (2012). Meiotic maps of sockeye salmon derived from massively parallel DNA sequencing. *BMC Genomics*, 13, 521.
- Ezaz, M. T., Harvey, S. C., Boonphakdee, C., Teale, A. J., McAndrew, B. J. & Penman, D. J. (2004). Isolation and physical mapping of sex-linked AFLP markers in Nile tilapia (*Oreochromis niloticus* L.). *Marine Biotechnology*, 6(5), 435–45.
- FAO (2016) The state of world fisheries and aquaculture. Report, 204.
- FAO, (2014). Global Aquaculture Production (FishStat), *Dicentrarchus labrax*.
- Ferraresso, S., Bonaldo, A., Parma, L., Cinotti, S., Massi, M., Bargelloni, L. and Gatta P.P. (2013). Exploring the larval transcriptome of the common sole (*Solea solea* L.). *BMC Genomics*, 14:315.
- Festing, M.F.W. (1992). The scope for improving the design of laboratory animal experiments, *Lab animals*, 26 (4): 256–268.
- Festing, M.F.W. (1995). Use of a Multistrain Assay Could Improve the NTP Carcinogenesis Bioassay. *Environmental Health Perspectives*, 103(1), 44–52.
- Festing, M.F.W. (1999). Warning : the use of heterogeneous mice may seriously damage your research. *Neurobiol. Aging*, 20 (2): 237–244.
- Festing, M.F.W. & Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. Institute for Laboratory Animal Research *ILAR Journal*, 43(4), 244–258.
- Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T., Young, G., Fennell, T., Allen, A., Ambrogio, L., Berlin, A.M., Blumenstiel, B., Cibulskis, K., Friedrich, D., Johnson, R., Juhn, F., Reilly, B., Shammass, R., Stalker, J., Sykes, S.M., Thompson, J., Walsh, J., Zimmer, A., Zwirko, Z., Gabriel, S., Nicol, R. & Nusbaum, C. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology*, 12:R1.
- Foisil, L. & Chourrout, D. (1992). Chromosome doubling by pressure treatment for tetraploidy and mitotic gynogenesis in rainbow trout *Oncorhynchus mykiss*: re-examination and improvements. *Aquaculture and Fisheries Management*, 23, 567–575.
- Francescon, A., Libertini, A., Bertotto, D. & Barbaro, A. (2004). Shock timing in mitogynogenesis and tetraploidization of the European sea bass *Dicentrarchus labrax*. *Aquaculture*, 236, 201–209.
- Francescon, A., Barbaro, A., Bertotto, D., Libertini, A., Cepollaro, F., Richard, J., Belvedere, P. & Colombo, L. (2005). Assessment of homozygosity and fertility in meiotic gynogens of the European sea bass (*Dicentrarchus labrax* L.). *Aquaculture*, 243(1-4), 93–102.
- Froese, R. & Pauly, D. (2014) FishBase. World Wide Web electronic publication.
- Fuji, K., Hasegawa, O., Honda, K., Kumasaka, K., Sakamoto, T., Okamoto, N. (2007). Marker assisted breeding of a lymphocystis disease-resistant Japanese flounder (*Paralichthys olivaceus*). *Aquaculture*, 272:291-295.

- Fujimura, K. & Okada, N. (2007). Development of the embryo, larva and early juvenile of Nile tilapia *Oreochromis niloticus* (Pisces: Cichlidae). *Development Growth Differentiation*, 49: 301–324.
- Galbusera, P., Volckaert, F. A. M. & Ollevier, F. (2000). Gynogenesis in the African catfish *Clarias gariepinus* (Burchell, 1822) III. Induction of endomitosis and the presence of residual genetic variation. *Aquaculture*, 185, 25–42.
- Garcia de Leon, F.J., Dallas, D.J., Chatain, B., Canonne, M., Versini, J.J. and Bonhomme, F. (1995). Development and use of microsatellite markers in seabass, *Dicentrarchus labrax* (Linnaeus, 1758) (Perciformes: Serranidae). *Molecular Marine Biology and Biotechnology*, 4, 62–68.
- Gharbi, K., Gautier, A., Danzmann, R. G., Gharbi, S., Sakamoto, T., Høyheim, B., Taggart, J.B., Caine, M., Powell, R., Krieg, F., Okamoto, N., Ferguson, M.M., Holm, L.E. & Guyomard, R. (2006). A linkage map for brown trout (*Salmo trutta*): chromosome homeologies and comparative genome organization with other salmonid fish. *Genetics*, 172(4), 2405–19.
- Gidskehaug, L., Kent, M., Hayes, B. J. & Lien, S. (2010). Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP-array. *Bioinformatics*, 1–9.
- Gjedrem, T. (1997). Flesh quality improvement in fish through breeding. *Aquaculture International*, (5): 197-206.
- Gjedrem, B. (2005) Design of Breeding Programs. In: Gjedrem, T., Ed., *Selection and Breeding Programs in Aquaculture*, Springer, Berlin, Heidelberg, 364.
- Gjedrem, T., Robinson, N. & Rye, M. (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture*, (350-353), 117-129.
- Gjedrem, T. & Robinson, N. (2014). Advances by selective breeding for aquatic species: a review. *Agricultural Sciences* 5:1152–1158.
- Glasauer, S. M. K. & Neuhauss, S. C. F. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics: MGG*, 289(6), 1045–60.
- Glencross, B. D., De Santis, C., Bicskei, B., Taggart, J. B., Bron, J. E., Betancor, M. B. & Tocher, D. R. (2015). A comparative analysis of the response of the hepatic transcriptome to dietary docosahexaenoic acid in Atlantic salmon (*Salmo salar*) post-smolts. *BMC Genomics*, 16, 684.
- Glover K. A., Hansen, M.M., Skaala, Ø. (2009). Identifying the source of farmed escaped Atlantic salmon (*Salmo salar*): Bayesian clustering analysis increases accuracy of assignment. *Aquaculture*, 290:37–46.
- Goldberg, A., Allis, C.D. & Bernstein, E. (2007). Epigenetics: A landscape takes shape. *Cell*, 23 (4): 635-638.
- Gomelsky, B. (2003). Chromosome set manipulation and sex control in common carp: a review. *Aquatic Living Resources*, 16(5), 408–415.
- Gonen, S., Lowe, N. R., Cezard, T., Gharbi, K., Bishop, S. C. & Houston, R. D. (2014). Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. *BMC Genomics*, 15, 166.
- Grant, B., Davie, A., Taggart, J. B., Selly, S. L. C., Picchi, N., Bradley, C., Prodohl, P., Leclercq, E. & Migaud, H. (2016). Seasonal changes in broodstock spawning

- performance and egg quality in ballan wrasse (*Labrus bergylta*). *Aquaculture*, 464, 505–514.
- Grimholt, U., Jonessen, R. & Smith, A.J. (2009). A review of the need and possible uses for genetically standardized Atlantic salmon (*Salmo salar*) in research. *Laboratory Animals*, 43, 121–126.
- Gu, X., Wang, Y. & Gu, J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet* 31: 205-209.
- Guo, X. & Allen, S. (1996). Complete Interference and Nonrandom Distribution Meiotic Crossover in a Mollusc, *Mulinia lateralis* (Say). *Biol Bull.*, 191(1), 145–148.
- Guy, D. R., Bishop, S. C., Brotherstone, S., Hamilton, A, Roberts, R. J., McAndrew, B. J. & Woolliams, J. A. (2006). Analysis of the incidence of infectious pancreatic necrosis mortality in pedigreed Atlantic salmon, *Salmo salar* L., populations. *Journal of Fish Diseases*, 29(11), 637–47.
- Guyomard, R., Mauger, S., Tabet-Canale, K., Martineau, S., Genet, C., Krieg, F. & Quillet, E. (2006). A type I and type II microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) with presumptive coverage of all chromosome arms. *BMC Genomics*, 7, 302.
- Guyon, R., Senger, F., Rakotomanga, M., Sadequi, N., Volckaert, F. A. M., Hitte, C. & Galibert, F. (2010). A radiation hybrid map of the European sea bass (*Dicentrarchus labrax*) based on 1581 markers: Synteny analysis with model fish genomes. *Genomics*, 96(4), 228–38.
- Hales, B.F., Grenier, L., Lalancette, C., Robaire, B. (2011). Epigenetic programming: from gametes to blastocyst. *Birth Defects Research Part A: Clinical and Molecular Teratology* 91, 652–665.
- Hall, A. T. (1999). BioEdit: A user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Oxford University Press. *Nucleic Acids symposium Series* N.41:96-98.
- Hamzah, A., Nguyen, H.N., Mekki, W., Ponzoni, R.W., Khaw, H, L., Abu Bakar, K.R. & Mohd Nor, S.A. (2016). Flesh characteristics: estimation of genetic parameters and correlated responses to selection for growth rate in the GIFT strain. *Aquaculture Research*, 47, 2139–2149.
- Han, H. S., Taniguchi, N. & Tsujimura, A. (1991). Production of Clonal Ayu by Chromosome Confirmation by Isozyme Marker and Manipulation and Tissue Grafting. *Nippon Suisan Gakkaishi*, 57(5), 825–832.
- Hara, M., Dewa, K., Yamamoto, E. (1993). DNA-fingerprinting with nonradioactive probe in clonal flounder *Paralichthys olivaceus*. *Nippon Suisan Gakkaishi*. 59:731–731.
- Hassanzadeh Saber, M. & Hallajian, A. (2013). Study of sex determination system in ship sturgeon, *Acipenser nudiventris* using meiotic gynogenesis. *Aquaculture International*, 22(1), 273–279.
- Haussler, D., O'Brien, S.J., Ryder, O.A., Barker, F.K., Clamp, M., Crawford, A.J., Hanner, R., Hanotte, O.H., Johnson, W., McGuire, J., Miller, W., Murphy, R.W., Murphy, W.J., Sheldon, F.H., Sinervo, B., Venkatesh, B., Wiley, E.O., Allendorf, F.W., Baker, S., Bernardi, G., Brenner, S., Cracraft, J., Diekhans, M., Edwards, S., Estes, J., Gaubert, P., Graphodatsky, A., Graves, J.A., Green, E.D., Hebert, P., Helgen, K.M., Kessing, B., Kingsley, D.M., Lewin, H.A., Luikart, G., Martelli, P., Nguyen, N., Orti, G., Pike, B.L., Rawson, D.M., Schuster, S.C., Seuánez, H.N.,

- Shaffer, H.B., Springer, M.S., Stuart, J.M., Teeling, E., Vrijenhoek, R.C., Ward, R.D., Wayne, R., Williams, T.M., Wolfe, N.D., Zhang, Y.P.(2009). A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *Journal of Heredity*. 2009, 100: 659-674.
- Havelka, M., Bytyutskyy, D., Symonová, R., Ráb, P., & Flajšhans, M. (2016). The second highest chromosome count among vertebrates is observed in cultured sturgeon and is associated with genome plasticity. *Genetics, Selection, Evolution : GSE*, 48, 12.
- Heston, W.E. (1968). Genetic aspects of experimental animals in cancer research. In: *Experimental animals in cancer research* (Muhlbock O, Nomura T, eds), Gann monograph 5. Tokyo: Maruzen, 1968; 3-15.
- Hoegg, S. & Meyer, A., 2005. Hox clusters as models for vertebrate genome evolution. *Trends Genet.* 21, 421–424.
- Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W. & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, 11 Suppl 1, 117–22.
- Hou, J., Saito, T., Fujimoto, T., Yamaha, E., & Arai, K. (2014). Androgenetic doubled haploids induced without irradiation of eggs in loach (*Misgurnus anguillicaudatus*). *Aquaculture*, 420-421, S57–S63.
- Hou, J., Fujimoto, T., Saito, T., Yamaha, E. & Arai, K. (2015). Generation of clonal zebrafish line by androgenesis without egg irradiation. *Nature Scientific Reports*, 5, 13346.
- Hou, J., Wang, G., Zhang, X., Sun, Z., Liu, H. & Wang, Y. (2016). Cold-shock induced androgenesis without egg irradiation and subsequent production of doubled haploids and a clonal line in Japanese flounder, *Paralichthys olivaceus*. *Aquaculture*, 464, 642–646.
- Houston, R.D., Cameron, N.D. & Rance, K.A. (2004). A melanocortin-4 receptor (MC4R) polymorphism is associated with performance traits in divergently selected large white pig populations. *Animal Genetics*, (35), 5, 386-390.
- Houston, R. D., Haley, C. S., Hamilton, A., Guy, D. R., Tinch, A. E., Taggart, J. B., McAndrew, B.J. & Bishop, S. C. (2008). Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*). *Genetics*, 178(2), 1109–15.
- Houston, R. D., Davey, J. W., Bishop, S. C., Lowe, N. R., Mota-Velasco, J. C., Hamilton, A., Guy, D. R., Tinch, A. E., Thomson, M. L., Blaxter, M. L., Gharbi, K., Bron, J. E. & Taggart, J. B. (2012). Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics*, 13, 244.
- Houston, R. D., Taggart, J. B., Cézard, T., Bekaert, M., Lowe, N. R., Downing, A., Talbot, R., Bishop, S.C., Archibald, A.L., Bron, J.E., Penman, D.J.P., Davassi, A., Brew F., Tinch, A.E., Gharbi, K. & Hamilton, A. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). *BMC Genomics*, 15, 90.
- Howe, K., Clark, M.D., Torroja, C.F., Tarrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assuncao, J.A., Zhou, Y., Gu, Y., Yen, J.,

- Vogel, J.H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G.K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Meida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Elliott, D., Threadgold, G., Harden, G., Ware, D., Mortimer, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C.M., Ersan-Urun, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberlander, M., Rudolph-Geiger, S., Teucke, M., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M., Enright, A., Geisler, R., Plasterk, R.H., Lee, C., Westerfield, M., De Jong, P.J., Zon, L., Postlethwait, J.H., Nusslein-Volhard, C., Hubbard, T.J., Roest, C.H., Rogers, J., Stemple, D.L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496:498–503.
- Hughes, A.L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99, 364–373.
- Hulata, G., A. Cnaani, N. Umiel, J.J. Agresti, S. Seki, B. May, S. Poompuang, E.M. Hallerman & G.A.E. Gall, (1999). Development of a tilapia artificial centre of origin and genetic linkage map based on AFLP and microsatellite loci, pp. 93–94 in *Towards Predictable Quality, Abstracts of Contributions Presented at the Aquaculture Europe 1999*. European Aquaculture Society Special Publication 27, Oostende, Belgium.
- Hulata, G. (2001). Genetic manipulations in aquaculture: a review of stock improvement by classical and modern technologies. *Genetica*, 111(1-3), 155–73.
- Hussain M.G., Penman D.J., McAndrew B.J. & Johnstone R. (1993) Suppression of first cleavage in the Nile tilapia, *Oreochromis niloticus* L. - a comparison of the relative effectiveness of pressure and heat shocks. *Aquaculture* 111, 263-270.
- Hussain, M. G., Penman, D. J. & McAndrew, B. J. (1998). Production of heterozygous and homozygous clones in Nile tilapia. *Aquaculture International*, (6) 197–205.
- Ihssen, P.E., Mckay, L. R., Mcmillan, I. & Phillips, R. B. (1990). Ploidy Manipulation and Gynogenesis in Fishes: Cytogenetic and Fisheries Applications. *Transactions of the American Fisheries Society*, 119, 698–717.
- Integrated DNA Technologies. Tips for using BLAST to locate PCR primers. (2015, June 23). Retrieved from <https://www.idtdna.com/pages/decoded/decoded-articles/pipet-tips/decoded/2011/03/16/tips-for-using-blast-to-locate-pcr-primers>.
- Jiang, L., Zhang, J., Wang, J.J., Wang, L., Zhang, L., Li, G., Yang, X., Ma, X., Sun, X., Cai, J., Zhang, J., Huang, X., Yu, M., Wang, X., Liu, F., Wu, C. I., He, C., Zhang,

- B., Ci,W., and Liu, J. ((2013). Sperm, but Not Oocyte, DNA Methylome Is Inherited by Zebrafish Early Embryos. *Cell*. 153(4): 773–784.
- Johnstone, R. (1992). Production and performance of triploid Atlantic salmon in Scotland. Scottish Aquaculture Research Report, No 2.
- Johnstone, R. & Stet, R. J. M. (1995). The production of gynogenetic Atlantic salmon, *Salmo salar* L. *Theoretical and Applied Genetics* 90: 6, 819-826.
- Johnson, I. A. (1999). Muscle development and growth: potential implications for flesh quality in fish. *Aquaculture*, (177): 99-115.
- Kai, W., Nomura, K., Fujiwara, A., Nakamura, Y., Yasuike, M., Ojima, N., Masaoka, T., Ozaki, A, Kazeto, Y., Gen, K., Nagao, J., Tanaka, H., Koyabashi, T. & Ototake, M. (2014). A ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (*Anguilla japonica*) provides insights into genome evolution after the teleost-specific genome duplication. *BMC Genomics*, 15, 233.
- Kakioka, R., Kokita, T., Kumada, H., Watanabe, K. & Okuda, N. (2013). A RAD-based linkage map and comparative genomics in the gudgeons (genus *Gnathopogon*, Cyprinidae). *BMC Genomics*, 14(1), 32.
- Kato, K., Hayashi, R., Yuasa, D., Yamamoto, S., Miyashita, S., Murata, O. & Kumai, H. (2002). Production of cloned red sea bream, *Pagrus major*, by chromosome manipulation. *Aquaculture*, 207(1-2), 19–27.
- Kent, M.P., Hayes, B., Xiang, Q., Berg, P.R., Gibbs, R.A. & Lien, S. (2009). Development of 16.5 K SNP chip for Atlantic Salmon. Proceedings of the 17th Plant and Animal Genome Conference, January 10–14, 2009. San Diego, CA.
- Khan, M., McAndrew, B. J. & Penman, D. J. (2014). Validation of clonal line females for sex determination studies in Nile tilapia *Oreochromis niloticus* L. *Research in Agriculture, Livestock and Fisheries*, 1(1), 147–158.
- Klasen, J. R., Piepho, H.P. & Stich, B. (2012). QTL detection power of multi-parental RIL populations in *Arabidopsis thaliana*. *Heredity*, 108(6), 626–32.
- Kobayashi, T., Ide, A., Hiasa, T., Fushiki, S., Ueno, K. (1994). Study of biological characters and the genetics of some traits of triploid and gynogenetic diploid on salmonid fishes: III. Production of cloned amago salmon *Oncorhynchus rhodurus*. *Fisheries Science*. 60:275–281.
- Kocher, T.D., Lee, W., Sobolewska, H., Penman, D. & MacAndrew, B. (1998). A Genetic linkage map of a chichlid fish the Tilapia (*O. niloticus*). *Genetics*, 148, 3, 1225-1232.
- Köhler, G. & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256 (5517):495–497.
- Komen, H. & Thorgaard G. (2007). Androgenesis , gynogenesis and the production of clones in fishes : A review. *Aquaculture*, 269, 150–173.
- Komen, J., Bongers, A.B.J., Richter, C.J.J., Van Muiswinkel, W.B. and Huisman, E.A. (1991) Gynogenesis in common carp (*Cyprinus carpio* L.) II. The production of homozygous gynogenetic clones and F1 hybrids. *Aquaculture* 92, 127–142.
- Komen, J., Boer, P. D. & Richter, C. J. J. (1992a). Male sex reversal in gynogenetic XX females of common carp (*C. carpio* L.) by a recessive mutation in a sex determining gene. *The Journal of Heredity*, 83, 431–434.
- Komen, J., Wiegertjes, G.F., Ginneken, V., Eding, E. & Richter, C.J.J. (1992b). Gynogenesis in common carp (*Cyprinus carpio*, L.). III.The effects of inbreeding

- on gonadal development of heterozygous and homozygous gynogenetic offspring. *Aquaculture*, 104, 51–66.
- Kurharczyk, D., Zarski, D., Targonska, K., Luczynski, M.J., Szczerboski, A., Nowosad, J., Kuwaja, R. & Mamcarz, A. (2014). Induced artificial androgenesis in common tench, *Tinca tinca* (L.), using common carp and common carp bream eggs. *Italian Journal of Animal Science*, 13, 196–200.
- Labbé, C., Robles, V. & Herraéz, M.P. (2016). Epigenetics in fish gametes and early embryo, *Aquaculture*, in press.
- Lahrech, Z., Kishioka, C., Morishima, K. & Mori, T. (2007). Genetic verification of induced gynogenesis and microsatellite – centromere mapping in the barfin flounder, *Verasper moseri*. *Aquaculture*, 272, S115–S124.
- Lal, M. M., Southgate, P. C., Jerry, D. R. & Zenger, K. R. (2016). Fishing for divergence in a sea of connectivity: The utility of ddRADseq genotyping in a marine invertebrate, the black-lip pearl oyster *Pinctada margaritifera*. *Marine Genomics*, 25, 57–68.
- Langmead, B. & Salzberg, S. (2013). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), 357–359.
- Larhammar, D. & Risinger, C. (1994). Molecular genetic aspects of tetraploidy in the common carp, *Cyprinus carpio*. *Mol Phylogenet Evol.* (3):59–68.
- Levasseur, A. & Pontarotti, P. (2011). The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biol. Direct* 6, 11.
- Li, R., Lyons, M. A., Wittenburg, H., Paigen, B. & Churchill, G. A. (2005). Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. *Genetics*, 169(3), 1699–709.
- Li, W.H. (1997). *Molecular Evolution*. Sinauer Associates, Sunderland.
- Lien, S., Gidskehaug, L., Moen, T., Hayes, B. J., Berg, P. R., Davidson, W. S., Omholt, S.W. & Kent, M. P. (2011). A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics*, 12, 615.
- Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., Schalburg, K.V., Rondeau, E.B., Genova, A.D., Samy, J.K.A., Vik, J.O., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Våge, D.I., Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A. J., Klunderud, A.T., Jakobsen, A.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S.W. & Davidson, W.S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, (6020).
- Limborg, M. T., Waples, R. K., Allendorf, F. W. & Seeb, J. E. (2015). Linkage Mapping Reveals Strong Chiasma Interference in Sockeye Salmon: Implications for Interpreting Genomic Data. *G3* (Bethesda, Md.), 5(11), 2463–73.
- Lin, F. & Dabrowski, K. (1998). Androgenesis and homozygous gynogenesis in muskellunge (*Esox masquinongy*): evaluation using flow cytometry. *Molecular Reproduction and Development* 49, 10–18.
- Liu, Y., Wang, G., Liu, Y., Hou, J., Wang, Y., Si, F. & Sun, Z. (2011). Production and verification of heterozygous clones in Japanese flounder, *Paralichthys olivaceus*

- by microsatellite marker. *African Journal of Biotechnology*, 10 (75), 17088–17094.
- Liu, Z.J. & Cordes, J.F. (2004). DNA marker technologies and their applications in aquaculture genetics. *Aquaculture* 238, 1–37.
- Liu, Z.J. (2011). *Next Generation Sequencing and Whole Genome Selection in Aquaculture*. Wiley-Blackwell.
- Locke, J. B., Aziz, R. K., Vicknair, M. R., Nizet, V. & Buchanan, J. T. (2008). Streptococcus iniae M-like protein contributes to virulence in fish and is a target for live attenuated vaccine development. *PLoS One*, 3(7), e2824.
- Lubieniecki, K. P., Jones, S. L., Davidson, E. a, Park, J., Koop, B. F., Walker, S. & Davidson, W. S. (2010). Comparative genomic analysis of Atlantic salmon, *Salmo salar*, from Europe and North America. *BMC Genetics*, 11, 105.
- Lubzens, E., Young, G., Bobe, J. & Cerdà, J. (2010). Oogenesis in teleosts: how eggs are formed. *General and Comparative Endocrinology*, 165(3), 367–89.
- Lucas, M. D., Drew, R. E., Wheeler, P. A., Verrell, P. A. & Thorgaard, G. H. (2004). Behavioral Differences Among Rainbow Trout Clonal Lines. *Behaviour Genetics*, 34(3): 355-365.
- Mable, B.K. (2004). “Why polyploidy is rarer in animals than in plants”: myths and mechanisms. *Biol. J. Linn. Soc.* 82, 453–466.
- Mair, G.C., J.S. Abucay, Beardmore, J.A. & Skibinski, D.O.F. (1995). Growth performance of genetically male tilapia (GMT) derived from YY-males in *Oreochromis niloticus* L.: on-station comparisons with mixed sex and sex reversed male populations. *Aquaculture* 137: 313–322.
- Manousaki, T., Tsakogiannis, A., Taggart, J. B., Palaiokostas, C., Tsaparis, D., Lagnel, J., Chatziplis, D., Magoulas, A., Papandroulakis, N., Mylonas, C. C. & Tsigenopoulos, C. S. (2015). Exploring a Non-model Teleost Genome Through RAD Sequencing-Linkage Mapping in Common Pandora, *Pagellus erythrinus* and Comparative Genomic Analysis. *G3* (Bethesda, Md.), 6(3), 509–19.
- Mardis, E. R. (2006). Anticipating the 1,000 dollar genome. *Genome Biology*, 7(7), 112.
- Margarido, G.R.A, Souza, A.P. & Garcia, A.A.F. (2007). OneMap: software for genetic mapping in outcrossing species. *Hereditas*, 144(3), 78–9.
- Martínez, P., Hermida, M., Pardo, B. G., Fernández, C., Castro, J., Cal, R. M., Alvarez-Dios, J.A., Gomez-Tato, A. & Bouza, C. (2008). Centromere-linkage in the turbot (*Scophthalmus maximus*) through half-tetrad analysis in diploid meiogynogenetics. *Aquaculture*, 280, 81–88.
- Martinez, V. (2007). Marker-assisted selection in fish and shellfish breeding schemes in Guimaraes, E., Ruane, J., Scherf, B.D., Sonniro, A. & Dargie, J.D. eds. *Marker Assisted Selection: Current status and future perspectives in crops, livestock, forestry and fish*. FAO, Viale delle Terme di Caracalla, 00153 Rome, Italy, pp. 329-363.
- May, B. & Grewe, P.M. (1993). Fate of maternal mtDNA following co-60 inactivation of maternal nuclear-DNA in unfertilized salmonid eggs. *Genome*. 36: 725–730.
- May, B. & Johnson, K. (1990) Composite linkage map of salmonid fishes. In: O’Brien SJ (ed) *Genetics maps*, 5th edn. Cold Spring Harbor, Cold Spring Harbor, 4151–4159.
- McCluskey, B. M. & Postlethwait, J. H. (2014). Phylogeny of zebrafish, a “model species,” within Danio, a “model genus”. *Molecular Biology and Evolution*, 32(3), 635–52.

- McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T. (2012). Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol* 66(2):526-538.
- Mesak, F., Tatarenkov, A., Earley, R. L. & Avise, J. C. (2014). Hundreds of SNPs vs. dozens of SSRs: which dataset better characterizes natural clonal lineages in a self-fertilizing fish? *Frontiers in Ecology and Evolution*, (2), 1–8.
- Migaud, H., Bell, G., Cbirita, E., McAndrew, B., Davie, A., Bobe, J., Herraiez, M.P. & Carrillo, M. (2013). Gamete quality and broodstock management in temperate fish. *Reviews in Aquaculture*, 5(Suppl.1), S194–S223.
- Milano, I., Babbucci, M., Panitz, F., Ogden, R., Nielsen, R. O., Taylor, M. I., Helyar, S.J., Carvalho, G.R., Espineira, M., Atanassova, M., Tinti, F., Maes, G. E., Patarnello, T., FishPopTrace Consortium and Bargelloni, L. (2011). Novel Tools for Conservation Genomics: Comparing Two High-Throughput Approaches for SNP Discovery in the Transcriptome of the European Hake. *Plos ONE*, 6(11): e28008.
- Millot, S., Péan, S., Labbé, L., Kerneis, T., Quillet, E., Dupont-Nivet, M. & Bégout, M.L. (2014). Assessment of genetic variability of fish personality traits using rainbow trout isogenic lines. *Behavior Genetics*, 44(4), 383–93.
- Mizgireuv, I. V & Revskoy, S. Y. (2006). Transplantable tumor lines generated in clonal zebrafish. *Cancer Research*, 66(6), 3120–5.
- Moen, T., Baranski, M., Sonesson, A. K. & Kjøglum, S. (2009). Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*): population-level associations between markers and trait. *BMC Genomics*, 10, 368.
- Morishima, K., Fujimoto, T., Sato, M., Kawae, A., Zhao, Y., Yamaha, E. & Arai, K. (2011). Cold-shock eliminates female nucleus in fertilized eggs to induce androgenesis in the loach (*Misgurnus anguillicaudatus*), a teleost fish. *BMC Biotechnology*, 11, 116.
- Morishima, K., Nakayama, I. & Arai, K. (2001). Microsatellite-centromere mapping in the loach, *Misgurnus anguillicaudatus*. *Genetica* (111), 59–69.
- Müller-Belecke, A. & Hörstgen-Schwark, G. (1995). Sex determination in tilapia (*Oreochromis niloticus*) sex ratios in homozygous gynogenetic progeny and their offspring. *Aquaculture* 137, 57–65.
- Müller-Belecke, A. & Hörstgen-Schwark, G. (2000). Performance testing of clonal *Oreochromis niloticus* lines. *Aquaculture*, 184, 67–76.
- Nam, Y.K., Chou, Y.S., Chou, J.C. & Kim D.S. (2000). Accelerated growth performance and stable germ-line transmission in androgenetically derived homozygous transgenic mud loach, *Misgurnus mizolepis*. *Aquaculture*, 209, 257-270.
- Naruse, K., Ijiri, K., Shima, A. & Egami N. (1985). The Production of Cloned Fish in the Medaka (*Oryzias latipes*). *J. Exp. Zool.* 236: 335-341.
- Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J. & Mitani, H. (2004). A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Research*, 14(5), 820–8.
- NC3Rs, National centre for the Replacement, Refinement & Reduction of animals in research. Retrieved October, 2016 from <https://www.nc3rs.org.uk/the-3rs>.
- Neira, R. (2010). Breeding in Aquaculture Species: Genetic Improvement Program in Developing Countries. 9th World Congress on genetics Applied to Livestock Production, Lipzig, 1-6 August 2010, 8.

- Nichols, K. M., Young, W. P., Danzmann, R. G., Robison, B. D., Rexroad, C., Noakes, M., Philips, R.B., Bentzen, P., Spies, I., Knudsen, K., Allendorf, F.W., Cunningham, B.M., Brunelli, B.M., Zhang, H., Ristow, S., Drew, R., Brown, K.H. Wheeler, P.A. & Thorgaard, G. H. (2003). A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). *Animal Genetics*, 34(2), 102–15.
- Nichols, K.M., Edo, A.F., Wheeler, P.A. & Thorgaard, G.H. (2008) The genetic basis of smoltification-related traits in *Oncorhynchus mykiss*. *Genetics*, 179, 1559–1575.
- Nie, H., Li, Q. & Kong, L. (2011). Microsatellite-centromere mapping in sea cucumber (*Apostichopus japonicus*) using gynogenetic diploid families. *Aquaculture*, 319(1-2), 67–71.
- Nie, H., Li, Q. & Kong, L. (2012). Centromere mapping in the Pacific abalone (*Haliotis discus hannai*) through half-tetrad analysis in gynogenetic diploid families. *Animal Genetics*, 43(3), 290–7.
- Nirea, K. G., Sonesson, A. K., Woolliams, J. A. and Meuwissen, T H.E. (2012). Strategies for implementing genomic selection in family-based aquaculture breeding schemes: double haploid sib test populations. *Genetics Selection Evolution*, 44:30.
- Nomura, K., Morishima, K., Tanaka, H., & Unuma, T. (2006). Microsatellite – centromere mapping in the Japanese eel (*Anguilla japonica*) by half-tetrad analysis using induced triploid families. *Aquaculture*, 257, 53–67.
- O’Neill, C.M., Morgan, C., Kirby, J., Tschöep, H., Deng, P. X., Brennan, M., Rosas, U., Fraser, F., Hall, C., Gill, S. & Bancroft, I. (2008). Six new recombinant inbred populations for the study of quantitative traits in *Arabidopsis thaliana*. *TAG. Theoretical and Applied Genetics*, 116(5), 623–34.
- Ocalewicz, K., Babiak, I., Dobosz, S., Nowaczyk, J. & Goryczko, K. (2004). The stability of telomereless chromosome fragments in adult androgenetic rainbow trout. *J Exp Biol*. 207: 2229–2236.
- Ocalewicz, K., Dobosz, S. & Kuzminski, H. (2012). Distribution of telomeric DNA sequences on the X-radiation-induced chromosome fragments observed in the genome of androgenetic brook trout (*Salvelinus fontinalis*, Mitchill 1814). *Cytogenetic and Genome Research*, 137(1), 1–6.
- Ohno, S., Wolf, U. & Atkin, N. (1967). Evolution from fish to mammals by gene duplication. *Hereditas*, 59(6).
- Ohno, S., 1970. *Evolution by Gene Duplication*. Springer, Berlin.
- Opazo, J. C., Butts, G. T., Nery, M. F., Storz, J. F. & Hoffmann, F. G. (2013). Whole-genome duplication and the functional diversification of teleost fish hemoglobins. *Molecular Biology and Evolution*, 30(1), 140–53.
- Ottera, H., Thorsen, A., Peruzzi, S., Dahle, G., Hansen, T. & Karlsen, Ø. (2011). Induction of meiotic gynogenesis in Atlantic cod, *Gadus morhua* (L.). *Journal of Applied Ichthyology*, 27, 1298–1302.
- Overturf. (2009). *Molecular Research in Aquaculture* (p. 395). Iowa 50014-8300, USA: Wiley-Blackwell.
- Ozaki, A., Sakamoto, T., Khoo, S., Nakamura, K., Coimbra, M.R.M., Akutsu, T. (2001). Quantitative trait loci (QTLs) associated with resistance/susceptibility to infectious pancreatic necrosis virus (IPNV) in rainbow trout (*Oncorhynchus mykiss*). *Molecular Genetics and Genomics*, 265(1), 23–31.
- Ozaki, A., Araki, K., Okamoto, H., Okauchi, M., Mushiake, K., Yoshida, K., Tsazaki, T., Fuji, K., Sakamoto, T. & Okamoto, N. (2012). Progress of DNA marker-assisted

- breeding in maricultured fin-fish. *Bulletin of Fisheries Research Agency* 35, 31–37.
- Palaiokostas, C., Bekaert, M., Davie, A., Cowan, M. E., Oral, M., Taggart, J. B., Gharbi, K., McAndrew, B.J., Penman, D.J. & Migaud, H. (2013a). Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing. *BMC Genomics*, 14, 566.
- Palaiokostas, C., Bekaert, M., Khan, M. G. Q., Taggart, J. B., Gharbi, K., McAndrew, B. J. & Penman, D. J. (2013b). Mapping and validation of the major sex-determining region in Nile tilapia (*Oreochromis niloticus* L.) Using RAD sequencing. *PLoS One*, 8(7), e68389.
- Palaiokostas, C., Bekaert, M., Khan, M. G. Q., Taggart, J. B., Gharbi, K., McAndrew, B. J. & Penman, D. J. (2015). A novel sex-determining QTL in Nile tilapia (*Oreochromis niloticus*). *BMC Genomics*, 16, 171.
- Palaiokostas, C., Bekaert, M., Taggart, J. B., Gharbi, K., McAndrew, B. J., Chatain, B., Penman, D.J. & Vandeputte, M. (2015). A new SNP-based vision of the genetics of sex determination in European sea bass (*Dicentrarchus labrax*). *Genetics Selection Evolution*, 47(1), 68.
- Pandian, T. J. & Koteeswaran, R. (1998). Ploidy induction and sex control in fish. *Hydrobiologia*, 384, 167–243.
- Pandian, T.J. & Kirankumar, S. (2003). Recent advances in hormonal induction of sex reversal in fish. *Journal of Applied Aquaculture*. 13: 205-230.
- Patton, S. J., Kane, S. L., Wheeler, P. A., & Thorgaard, G. H. (2007). Maternal and paternal influence on early embryonic survival of androgenetic rainbow trout (*Oncorhynchus mykiss*): Implications for measuring egg quality. *Aquaculture*, 263(1-4), 26–34.
- Pecoraro, C., Babbucci, M., Villamor, A., Franch, R., Papetti, C., Leroy, B., Garcia, S. O., Muir, J., Rooker, J., Arocha, F., Murua, H., Zudaire, I., Chassot, E., Bodin, N., Tinti, F., Bargelloni, L. and Cariani, A. (2016). Methodological assessment of 2b-RAD genotyping technique for population structure inferences in yellowfin tuna (*Thunnus albacares*). *Marine Genomics* 25; 43–48.
- Peer, Y.V., Maere, S. & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* 10, 725-732.
- Penman, D. J. & Piferrer, F. (2008). Fish Gonadogenesis. Part I: Genetic and Environmental Mechanisms of Sex Determination. *Reviews in Fisheries Science*, 16(sup1), 16–34.
- Penman, D.J. & McAndrew, B.J. (2000). Genetics for the management and improvement of cultured tilapia in *Tilapias: Biology and Exploitation* (Eds. Beveridge, C.M. & McAndrews, J.B.) (pp.227-255). Springer Academic Publishers.
- Peruzzi, S. & Chatain, B. (2000). Pressure and cold shock induction of meiotic gynogenesis and triploidy in the European sea bass, *Dicentrarchus labrax* L.: relative efficiency of methods and parental variability. *Aquaculture*, 189, 23–37.
- Peterson B.K., Weber, J., Kay E.H., Fisher, H. S. & Hoekstra, H.E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS One*, 7(5), e37135.
- Piferrer, F., Benfey, T.J., Donaldson, E.M. (1994). Gonadal morphology of normal and sex-reversed triploid and gynogenetic diploid coho salmon (*Oncorhynchus kisutch*). *Journal of Fish Biology*. 45, 541–553.

- Piferrer, F. (2013). Epigenetics of sex determination and gonadogenesis. *Developmental dynamics*, Special issue reviews-A, 242, 360-370.
- Pongthana, N., Penman, D. J., Baoprasertkul, P., Hussain, M. G., Islam, M. S., Powell, S. F. & McAndrew, B. J. (1999). Monosex female production in the silver barb (*Puntius gonionotus*, Bleeker). *Aquaculture*, 173, (1-4), 247-256.
- Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Research* (20), 291–300.
- Potok, M. E., Nix, D.A., Parnell, T. J. and Bradley R. Cairns, B. R. (2013). Reprogramming the Maternal Zebrafish Genome after Fertilization to Match the Paternal Methylation Pattern. *Cell* 153, 759–772.
- Preston, A. C., Taylor, J. F., Craig, B., Bozzolla, P., Penman, D. J. & Migaud, H. (2013). Optimisation of triploidy induction in brown trout (*Salmo trutta* L.). *Aquaculture*, 414, 160–166.
- Purcell, M. K., Nichols, K. M., Winton, J. R., Kurath, G., Thorgaard, G. H., Wheeler, P., Hansen, J.D., Herwih, R.P. & Park, L.K. (2006). Comprehensive gene expression profiling following DNA vaccination of rainbow trout against infectious hematopoietic necrosis virus. *Molecular Immunology*, 43(13), 2089–106.
- Purdom, C.E. (1983). Genetic engineering by the manipulations of chromosomes. *Aquaculture*, 33, 287–300.
- Purdom, C.E. (1993). *Genetics and Fish Breeding*. Chapman and Hall, London.
- Puritz, J. B., Matz, M. V, Toonen, R. J., Weber, J. N., Bolnick, D. I. & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology*, 23(24), 5937–42.
- Qin, Q.W., Ototake, M., Nagoya, H., Nakanishi, T. (2002). Graft-versus-host reaction(GVHR) in clonal amago salmon, *Oncorhynchus rhodurus*. *Veterinary Immunology and Immunopathology*. 89:83–89.
- Quillet, E., Garcia, P. & Guyomard, R. (1991). Analysis of the Production of All Homozygous Lines of Rainbow Trout by Gynogenesis. *The Journal of Experimental Zoology*, 374, 367–374.
- Quillet, E. (1994). Survival, growth and reproductive traits of mitotic gynogenetic rainbow trout females. *Aquaculture*, 123(3-4), 223–236.
- Quillet, E., Aubard, G. & Quéau, I. (2002). Mutation in a sex-determining gene in rainbow trout: detection and genetic analysis. *The Journal of Heredity*, 93(2), 91–9.
- Quillet, E., Dorson, M., Le Guillou, S., Benmansour, A. & Boudinot, P. (2007). Wide range of susceptibility to rhabdoviruses in homozygous clones of rainbow trout. *Fish & Shellfish Immunology*, 22(5), 510–9.
- Recknagel, H., Elmer, K. R. & Meyer, A. (2013). A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3 (Genes, Genomes, Genetics)*, 3(1), 65–74.
- Reddy, P.V.G.K. (1999). Genetic resources of Indian major carps. FAO Fish. Tech. Pap. Page: 387-396.
- Reid, D. P., Smith, C. A., Rommens, M., Blanchard, B., Martin-Robichaud, D. & Reith, M. (2007). A Genetic linkage map of Atlantic halibut (*Hippoglossus hippoglossus* L.). *Genetics*, 177(2), 1193–205.
- Robison, B.D., Wheeler, P.A. & Thorgaard, G.H. (1999). Variation in development rate among clonal lines of rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, 173, 131–141.

- Robison, B.D. & Thorgaard, G.H. (2011) Prospects and pitfalls of clonal fishes in the postgenomic era in *Aquaculture Biotechnology* (Ed.Fletcher, G.L. & Rise, M.R.) (pp.166-192). Wiley-Blackwell publishing, Chichester, West Sussex, UK.
- Rokas, A. & Abbot, P. (2009). Harnessing genomics for evolutionary insights. *Trends in Ecology & Evolution*, 24(4), 192–200.
- Rondeau, E. B., Minkley, D. R., Leong, J. S., Messmer, A. M., Jantzen, J. R., von Schalburg, K. R., Lemon, C., Bird, N.H. & Koop, B. F. (2014). The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. *PloS One*, 9(7).
- Rye, M., Gjerde, B. & Gjedrem, T. (2010). Genetic Development Programs for Aquaculture Species in Developed Countries. 9th World Congress on Genetics Applied to Livestock Production, Lipzig, 1-6 August 2010, 8.
- Saillant, E., Fostier, A., Haffray, P., Menu, B., Thimonier, J. & Chatain, B. (2002). Temperature effects and genotype-temperature interactions on sex determination in the European sea bass (*Dicentrarchus labrax* L.). *The Journal of Experimental Zoology*, 292(5), 494–505.
- Sakamoto, T., Danzmann, R. G., Gharbi, K., Howard, P., Ozaki, A., Khoo, S. K., Woram, R.A., Okamoto, N., Ferguson, M.M., Holm, L.E., Guyomard, R. & Hoyheim, B. (2000). A Microsatellite Linkage Map of Rainbow Trout (*Oncorhynchus mykiss*) Characterized by Large Sex-Specific Differences in Recombination Rates. *Genetics Society of America*, 155 (3): 1331–1345.
- Sánchez et al., (2011). SNP Analysis with Duplicated Fish Genomes: Differentiation of SNPs, Paralogous Sequence Variants, and Multisite Variants in Liu, Z.J. (ed.), *Next Generation Sequencing and Whole Genome Selection in Aquaculture* (133-150 pg), Blackwell Publishing Ltd.
- Sarder, M. R., Penman, D. J., Myers, J. M. & McAndrew, B. J. (1999). Production and propagation of fully inbred clonal lines in the Nile tilapia (*Oreochromis niloticus* L.). *The Journal of Experimental Zoology*, 284(6), 675–685.
- Sato, Y. & Nishida, M., 2010. Teleost fish with specific genome duplication as unique models of vertebrate evolution. *Environ. Biol. Fishes* 88, 169–188.
- Seoighe, C. & Wolfe, K.H., (1999). Updated map of duplicated regions in the yeast genome. *Gene* 238, 253–261.
- Smedley, M. A., Clokie, B. G. J., Migaud, H., Campbell, P., Walton, J., Hunter, D., Corrigan, D. & Taylor, J. F. (2016). Dietary phosphorous and protein supplementation enhances seawater growth and reduces severity of vertebral malformation in triploid Atlantic salmon (*Salmo salar* L.). *Aquaculture*, 451, 357–368.
- Smith, C.T., Elfstrom, C.M., Seeb, L. W. & Seeb, J. E. (2005). Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology*, 14(13), 4193–203.
- Smith, M., Asche, F., Guttormsen, A. G. & Wiener, B.J. (2010). Genetically Modified Salmon and full impact assesment. *Science*, 330, 1052–1053.
- Smith, J.J., Kuraku, S., Holt, C., Spengler, T., Jiang, N., Campbell, M.S., Yandell, D., Manusaki, T., Meyer, A., Bloom, O.E., Morgan, J.R., Buxbaum, J.D., Sachidanandam, R. et al. (2013). Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genetics*, 45, 415–421.

- Snell, G.D., Dausset, J. & Nathenson, S.G. (1976). *Histocompatibility*. New York: Academic, Print.
- Sola, L., Bressanello, S., Rossi, A., Iaselli, V., Crosetti, D. & Cataudella, S. (1993). A karyotype analysis of the genus *Dicentrarchus* by different staining techniques. *Journal of Fish Biology*, 43, 329–337.
- Sonesson, A. (2011). Genomic selection for aquaculture: principles and procedures. In Liu, Z. J. Eds. *Next Generation Sequencing and Whole Genome Selection in Aquaculture*. Iowa, USA, Blackwell publishing, pp. 151-163.
- Streisinger, C., Walkern, C., Dowerd, N., Knauber, D. & Singer, F. (1981) Production of clones of homozygous diploid zebrafish (*Brachydanio rerio*). *Nature* 291: 293-296.
- Sun, Y., Zhang, C., Liu, S., Duan, W., Liu, Y. (2007). Induced interspecific androgenesis using diploid sperm from allotetraploid hybrids of common carp × red crucian carp. *Aquaculture* 264: (1–4), 47–53.
- Suquet, M.; Omnes, M.H.; Normant, Y.; Fauvel C. (1992). Assessment of sperm concentration and motility in turbot (*Scophthalmus maximus*). *Aquaculture*, 101, 177-185.
- Tanck, M. W., Claes, T., Bovenhuis, H. & Komen, J. (2002). Exploring the genetic background of stress using isogenic progenies of common carp selected for high or low stress-related cortisol response. *Aquaculture*, 204:(3-4), 419–434.
- Taniguchi, N. T., Hana, H.S. & Tsujimura, A. (1994). Variation in some quantitative traits of clones produced by chromosome manipulation in ayu, *Plecoglossus altivelis*. *Aquaculture*, 120, 53–60.
- Taniguchi, N., Yamasaki, M., Takagi, M., Tsujimura, A. (1996). Genetic and environmental variances of body size and morphological traits in communally reared clonal lines from gynogenetic diploid ayu, *Plecoglossus altivelis*. *Aquaculture*. 140:333–341.
- Taslina, K., Davie, A., McAndrew, B. J. & Penman, D. J. (2015). DNA sampling from mucus in the Nile tilapia, *Oreochromis niloticus*: minimally invasive sampling for aquaculture-related genetics research. *Aquaculture Research*, 1-6.
- Thorgaard, G.H., Bailey, G.S., Williams, D., Buhler, D.R., Kaattari, S.L., Ristow, S.S., Hansen, J.D., Winton, J.R., Bartolomew, J.L., Nagler, J.J., Walsh, P.J., Vijayan, M.M., Devlin, R.H., Hardy, R.W., Overtuf, K.E., Young, W.P., Robison, B.D., Rexford, C. & Palti, Y. (2002) Status and opportunities for genomics research with rainbow trout. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology*, 133, 609–646.
- Thorgaard, G.H. (1983). Gene-Centromere mapping in rainbow trout: High interference over long map distances. *Genetics*, 103, 771–783.
- Thorgaard, G.H. (1986). Ploidy manipulation and performance. *Aquaculture* 57:57–64
- Timusk, E. R., Ferguson, M. M., Moghadam, H. K., Norman, J. D., Wilson, C. C. & Danzmann, R. G. (2011). Genome evolution in the fish family salmonidae: generation of a brook charr genetic map and comparisons among charrs (Arctic charr and brook charr) with rainbow trout. *BMC Genetics*, 12(1), 68.
- Tine, M., Kuhl, H., Gagnaire, P.A., Louro, B., Desmarais, E., Martins, R. S. T., Hecht, J., Knaust, F., Belkhir, K., Klages, S., Dieterich, R., Stueber, K., Piferrer, F., Guinand, B., Bierne, N., Volckaert, F. A. M., Bargelloni, L., Powe, D. M., Bonhomme, F., Canario, A. V. M. & Reinhardt, R. (2014). European sea bass

- genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, 5, 5770.
- Tvedt, H.B., Benfey, T.J., Martin-Robichaud, D.J., McGowan, C. & Reith, M. (2006). Gynogenesis and sex determination in Atlantic Halibut (*Hippoglossus hippoglossus*). *Aquaculture*, 252(2–4), 573–583.
- Vandeputte, M., Dupont-Nivet, M., Chavanne, H. & Chatain, B. (2007). A polygenic hypothesis for sex determination in the European sea bass *Dicentrarchus labrax*. *Genetics*, 176(2), 1049–57.
- Vandeputte, M., Dupont-Nivet, M., Haffray, P., Chavanne, H., Cenadelli, S., Parati, K., Vidal, M. O., Vergnet, A. and Chatain, B. (2009). Response to domestication and selection for growth in the European sea bass (*Dicentrarchus labrax*) in separate and mixed tanks. *Aquaculture*, 286(1-2), 20–27.
- Varadaraj, K., (1993). Production of viable haploid *Oreochromis mossambicus* gynogens using UV-irradiated sperm. *J. Exp. Zool.* 267, 460–467.
- Vasemägi, A., Nilsson, J. & Primmer, C. R. (2005). Seventy-five EST-linked Atlantic salmon (*Salmo salar* L.) microsatellite markers and their cross-amplification in five salmonid species. *Molecular Ecology Notes*, 5(2), 282–288.
- Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., Swann, J.B., Ohta, Y., Flajnik, M., Sutoh, Y., Kasahara, M., Hoon, S., Gangu, V., Roy, S.W., Irimia, M., Korzh, V., Kondrychyn, I., Lim, Z.W., Tay, B.H. et al. (2014). Elephant shark genome provides unique insights into gnathostome evolution. *Nature* 505, 174–179.
- Wei, C.M., Li, J.N. & Bumgarner, R.E. (2004). Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*. 5: 87.
- Wolfe, K.H., (2001). Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341.
- Wright, J.E., Johnson, K., Hollister, A. & May, B. (1983). Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. *Isozymes Curr. Top. Biol. Med. Res.* 10: 239–260.
- Wu, R., Xing Ma, C., Wu, S. S. & Zeng, Z. (2002). Linkage mapping of sex-specific differences. *Genetic Research, Cambridge Journal*, 79, 85–96.
- Xu, P. Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., Zhang, Y., Hou, L., Zhao, Z., Cao, D., Lu, C., Li, C., Zhou, Y., Liu, Z., Fan, Z., Shan, G., Li, X., Wu, S., Song, L., Hou, G., Jiang, Y., Jeney, Z., Yu, D., Wang, L., Shou, C., Song, L., Sun, J., Li, P., et al. (2014). Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nature Genetics*, 46, 11.
- Yamaha, E., Otani, S., Minami, A. & Arai, K. (2002). Dorso-ventral axis perturbation in goldfish embryos caused by heat and pressure shock treatments for chromosome set manipulation. *Fisheries Science* 68, 313–319.
- Yamamoto, E. (1999). Studies on sex-manipulation and production of cloned populations in hirame, *Paralichthys olivaceus* (Temminck et Schlegel). *Aquaculture*, 173, 235–246.
- Young, W.P., Wheeler, P.A., Fields, R.D., Thorgaard, G.H. (1996) DNA fingerprinting confirms isogenicity of androgenetically derived rainbow trout lines. *The Journal of Heredity*, 87, 77–80.

- Young, W., Wheeler, P., Coryell, V., Keim, P. & Thorgaard, G. (1998). A Detailed Linkage Map of Rainbow Trout Produced Using Doubled Haploids. *Genetics*, 148, 839–850.
- Yue, G. H. (2013). Recent advances of genome mapping and marker-assisted selection in aquaculture. In P. H. and G. C. Tony Pitcher (Ed.), *Fish and Fisheries*. John Wiley & Sons Ltd.
- Zhang, H., Tan, E., Suzuki, Y., Hirose, Y., Kinoshita, S., Okano, H., Kudoh, J., Shimizu, A., Saito, K., Watabe, S. and Asakawa, S. (2014). Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Scientific reports* 4, 6780.
- Zhou, R., Xiao, J., Qin, Q., Zhu, B., Zhao, R., Zhang, C., Tao, M., Luo, K., Wang, J., Peng, L. & Liu, S. (2015). YY super sperm lead to all male triploids and tetraploids. *BMC Genetics*, 16:68.
- Zhu, C., Sun, Y., Yu, X. & Tong, J. (2013). Centromere Localization for Bighead Carp (*Aristichthys nobilis*) through Half-Tetrad Analysis in Diploid Gynogenetic Families. *PLoS One*, 8(12), e82950.
- Zimmerman, A. M., Evenhuis, J. P., Thorgaard, G. H. & Ristow, S. S. (2004). A single major chromosomal region controls natural killer cell-like activity in rainbow trout. *Immunogenetics*, 55(12), 825–35.
- Zimmerman, A.M., Wheeler, P.A., Ristow, S.S. & Thorgaard, G.H. (2005). Composite interval mapping reveals three QTL associated with pyloric caeca number in rainbow trout, *Oncorhynchus mykiss*. *Aquaculture*, 247 : 85 - 95 .
- Zou, F., Gelfond, J. A. L., Airey, D. C., Lu, L., Manly, K. F., Williams, R. W. & Threadgill, D. W. (2005). Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. *Genetics*, 170(3), 1299–311.

Appendix

Chapter 2.1

SSTNE buffer receipt for salt precipitation DNA extraction protocol: for 1 L

17.5 g NaCl

6.05 g Tris Base

1 mL EDTA 0.2 M

76 mg EGTA (E3889 Sigma Aldrich)

72 mg spermidine (S0266, Sigma Aldrich)

52 mg spermine (S1141, Sigma Aldrich)

Autoclave and store at 4°C

pH is c. 9.5 -10.0

Do not require pH adjustment.

Do not require vortex at any stage of preparation.

Preparation volume can be scaled up as required.

Chapter 2.2

Microsatellite markers used in Chapter 5 (Atlantic salmon). Source for all microsatellites is (Vasemägi, Nilsson, & Primmer, 2005). Size standard 600bp was used for multiplex panel_1, while multiplex panel_2 and 3 used SS400bp.

Locus name	Repeat	Alleles	Primer 5'--> 3'	Panel	Dye	Size range (bp)	Ta (°C)
BG935488	(CAAT) ₂₃	18	F: TGACCCACCAAGTTTTTCT R: GTTTAAACACAGTAAGCCCATCTATTG	1	M13A_blue	166–234	60
CA048828	(CA) ₁₉	24	F: GAGGGCTTCCCATAACAACAA R:GTTTAAGCGGTGAGTTGACGAGAG	2	M13A_blue	251–307	60
CA060177	(TGAG) ₁₈	27	F: CGCTTCCTGGACAAAAATTA R: GTTTGAGCACACCCATTCTCA	3	M13A_blue	294–374	60
CA038592	(AT) ₁₂	24	F: AAGCATCAAACCAACCTCATT R: GTTTCGGGGGTGAAGATGTCTACT	1	M13A_blue	340–426	60
CA055301	(CA) ₂₉	21	F: AGAACCAAGGGTACCGATCC R: GTTTGGGAAATGGGTGGTAAGAAAA	2	CAG_green	217–263	60
CA053480	(AC) ₁₅	16	F: TGGTCACAAACCAAATGGAA R: GTTCCACTCCAGGGTGTCTGTAA	1	CAG_green	254–290	60
CB515794	(GT) ₂₆	16	F: CTCAGTGCCATGTCTCCAAC R: GTTTCATCCTGTCTGCTGACTG	3	CAG_green	265–309	60
CA059136	(TA) ₂₂	27	F: AGGGTAGTGAGAAAGCAGCAA R: GTTTAACTGGCTGGCCATAGG	2	CAG_green	318–380	60
BG934281	(TCTG) ₁₄	27	F: ACTGCTTCTCCCCTGCTACA R: GTTTGCGAACCACACATATACCAC	2	Goddle_black	193–267	60
CA048302	(AC) ₂₀	23	F:TTGCCACCTCTAAACGCTTC R:GTTTAAATGAACCCAGCCATACA	3	Goddle_black	201–255	60
CB517044	(TA) ₂₁	28	F: CACCAAGCATGGGAAGCTAT R: GTTTGCTGCCACACAGGCTACTTT	1	Goddle_black	351–425	55

Chapter 2.3

Modified Fish Ringers (MFR) solution receipt: for 500 ml

3.25 g NaCl

2.50 g KCl

0.10 g NaHCO₃

0.15 g CaCl₂, 6H₂O

72 mg spermidine (S0266, Sigma Aldrich)

52 mg spermine (S1141, Sigma Aldrich)

pH is c. 8.3

Store at +4°C

Does not require autoclaving.

Preparation volume can be scaled up as required.

- Chapter 3.1_Report S1:** A comprehensive quality report, FASTQC, on raw reads of P1*
- Chapter 3.2_Report S2:** A comprehensive quality report, FASTQC, on raw reads of P2*
- Chapter 3.3_Table S1:** Detailed information for each sample used: Sample ID, origin, UV irradiation and shock parameters, sampling tissue, fertilisation and sampling date*
- Chapter 3.4_Table S2:** ddRAD runs comparison (1st, 2nd and 1st+2nd sequencing run)*
- Chapter 3.5_Table S3:** All SNP markers used: Marker ID, locations of markers on physical map, genetic map corresponding of LGs, distance (cM) and the percentage of heterozygosity ratio. (*: marker that have been assigned in to one big LG and this linkage could not been broken with increasing LOD scores)*
- Chapter 3.6_Table S4:** Position of microsatellites from previous studies and their informative level with frequency distribution*
- Chapter 3.7_Table S5:** Crossover points per chromosome arm*
- Chapter 3.8_Table S6:** Marker-centromere recombination rate (y) and map distances of 804 female heterogametic loci examined in meiotic gynogenetic seabass family*
- Chapter 3.9_Dataset S1:** Marker ID, physical map location, percentage recombination frequency and sequences. (FASTA format)*
- Chapter 4.1_Report S1:** A comprehensive quality report, FastQC, on raw reads of P1*
- Chapter 4.2_Report S2:** A comprehensive quality report, FastQC, on raw reads of P2*
- Chapter 4.3_Dataset S1:** The sequences of the markers used for the verification of isogenic clone founders in each family*
- Chapter 4.4_Table S1:** A detailed information for each samples used: Sample ID, description, family, gender, family origin, repeat level in the ddRAD library and P1 and P2 barcodes used per sample*
- Chapter 4.5_Table S2:** A detailed association graphs on a single meiotic gynogenetic identified in F6 family. The data from previous chapter-3, meiotic gynogenetics, were used to plot % heterozygosity versus genome assembly (Mbp) in each linkage group as well as the dataset produced in the present study *
- Chapter 5.1_Report S1:** A comprehensive quality report, FastQC, on raw reads of P1*
- Chapter 5.2_Report S2:** A comprehensive quality report, FastQC, on raw reads of P2*
- Chapter 5.3_Dataset S1:** Sequences of all one-copy markers used for verification study*
- Chapter 5.4_Material S1:** BLAST scrips for removing multi-copy loci*
- Chapter 5.5_Table S1:** A detailed information for each samples used: Sample ID, type, description, family, gender, repeat and barcodes (P1 & P2)*
- Chapter 5.6_Table S2:** Physical position of microsatellites used*
- Chapter 6.1_Report S1:** A comprehensive quality report, FastQC, on raw reads of P1*
- Chapter 6.2_Report S2:** A comprehensive quality report, FastQC, on raw reads of P2*

*: Available in electronic version.