UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Biomedicina Comparata e Alimentazione

Corso di laurea magistrale in Biotecnologie per l'alimentazione

La genetica della determinazione del sesso nel cefalo comune (*Mugil cephalus* Linnaeus, 1758): analisi di polimorfismi nucleotidici mediante whole genome sequencing

Relatore
Prof. Luca Bargelloni

Correlatore
Dr.ssa Serena Ferraresso

Laureanda
Elisa Corsini
Matricola n.
2052460

ANNO ACCADEMICO 2022/2023

Riassunto

Il cefalo (*Mugil cephalus*), la specie più diffusa della famiglia Mugilidae, è un animale cosmopolita che vive nelle acque costiere dei principali oceani delle zone tropicali, subtropicali e temperate. L'Egitto è il più grande produttore di cefalo d'allevamento. Anche la Repubblica di Corea, l'Italia, la provincia cinese di Taiwan e Israele sono produttori importanti.

La bottarga di muggine è un pregiato prodotto alimentare che si ottiene da un processo di salagione ed essiccazione delle uova di *M. cephalus* e può raggiungere un prezzo di 300 €/Kg. In Italia la regione più importante per la produzione di bottarga di muggine è la Sardegna. L'elevato valore commerciale di questo prodotto, e di conseguenza delle femmine di *M. cephalus*, ha portato ad un crescente interesse nei confronti dei meccanismi genetici di determinazione del sesso in questa specie. La capacità di controllare la sex ratio, con una produzione quasi esclusivamente femminile, e/o l'identificazione di marcatori genetici in grado di sessare precocemente gli individui porterebbero enormi vantaggi all'industria del cefalo.

I pesci teleostei presentano una grande varietà di meccanismi di determinazione del sesso che comprendono fattori genetici (GSD, *Genetic Sex Determination*), ambientali (ESD, *Environmental Sex Determination*) e sociali. *M. cephalus* è una specie gonocorica, ma i meccanismi che regolano la sua determinazione del sesso non sono ancora chiari. Sono stati condotti diversi studi con lo scopo di individuare possibili geni che permettano di effettuare una precoce determinazione del sesso.

In uno studio condotto da Dor et al. sono stati individuati sette geni potenzialmente associati alla determinazione del sesso: *gth-ri* (*fshr*), *foxi1*, *dhx32*, *bub3*, *dock1*, *bccip* e *dhx32a*. In un successivo lavoro condotto da Ferraresso et al. nel 2021 sono stati analizzati attraverso Pool-Sequencing individui di *M. cephalus* appartenenti a diverse popolazioni. Lo studio ha permesso di identificare tre polimorfismi a singolo nucleotide (SNP) sull'esone 14 del gene *fshr* che potrebbero contribuire alla determinazione del sesso. Queste mutazioni sono significativamente associate al sesso fenotipico, anche se un certo numero di maschi, definiti "non conformi" hanno mostrato un genotipo wt/wt, normalmente associato al fenotipo femminile.

Un diverso patrimonio genetico, come anche diverse condizioni ambientali, possono giocare un ruolo importante andando a modulare l'azione di *fshr*. In questo contesto, l'obiettivo del presente lavoro di tesi è stato quello di valutare l'esistenza di altre mutazioni capaci di influenzare la determinazione del sesso in *M. cephalus*. Per raggiungere tale scopo, il genoma di 32 individui di sesso femminile di cefalo,

appartenenti a due diverse popolazioni, è stato sequenziato con tecnologia Illumina e confrontato con il genoma di 40 maschi conformi e non conformi appartenenti alle stesse popolazioni geografiche.

Abstract

The flathead grey mullet (*Mugil cephalus*), the most widespread species of the family Mugilidae, is a cosmopolitan animal that lives in waters around the coast of the major oceans of tropical, subtropical and temperate zones. Egypt is the largest producer of farmed mullet, followed by Republic of Korea, Italy, Taiwan and Israel.

Mullet roe is a valuable food product obtained by a salting and drying process of *M. cephalus* eggs and it can reach a price of 300 €/kg. In Italy, Sardinia is the most important region for the mullet roe production. The high commercial value of this product, and consequently of *M. cephalus* females, has led to a growing interest in the genetic sex determination mechanisms in this species. The ability to control the sex ratio, with almost exclusively female production, or the identification of genetic markers able to sexing individuals early would bring enormous advantages to the mullet industry.

A large variety of sex determination mechanisms characterize teleost fish, including genetic (GSD, Genetic Sex Determination), environmental (ESD, Environmental Sex Determination) and social factors. *M. cephalus* is a gonochoric species but the mechanisms regulating its sex determination are still unclear. Several studies have been conducted with the aim of identifying possible genes that allow early sex determination.

In a study conducted by Dor et al., seven genes potentially associated with sex determination were identified: *gth-ri* (also known as *fshr*), *foxi1*, *dhx32*, *bub3*, *dock1*, *bccip* and *dhx32a*. In a work conducted by Ferraresso et al. in 2021, mullet from different populations were analyzed by Pool-Sequencing. The study allowed to identify three single nucleotide polymorphisms (SNPs) on exon 14 of the *fshr* gene that could contribute to sex determination. These mutations are significantly associated with phenotypic sex, although some males, defined as “non-conformi”, showed a wt/wt genotype, normally associated with the female phenotype.

A different genetic heritage, as well as different environmental conditions, can play an important role by modulating *fshr* activity. In this context, the purpose of this thesis work was to evaluate the existence of other mutations able to influence sex determination in *M. cephalus*. To achieve this aim, the genome of 32 female mullet individuals, belonging to two different populations, was sequenced with Illumina technology and compared with the genome of 40 compliant and non-compliant males belonging to the same geographical populations.

Indice

1. Introduzione

1.1. La determinazione del sesso nei vertebrati

Gran parte degli animali appartenenti al gruppo dei vertebrati presenta sessi separati (gonocorismo) e la distinzione tra individui di sesso maschile e femminile avviene principalmente attraverso due strategie:

- Determinazione genetica del sesso (GSD)
  La determinazione del sesso è guidata da elementi genetici e il sesso di un individuo è solitamente definito al momento della formazione dello zigote. I due sistemi più comuni per la determinazione del sesso sono il sistema eterogametico maschile con cromosomi sessuali XY e il sistema eterogametico femminile con cromosomi ZW.
- Determinazione ambientale del sesso (ESD)
  La determinazione del sesso viene attivata dagli effetti di fattori ambientali durante lo sviluppo. Alcuni stimoli possono essere: temperatura, fotoperiodo (ore di luce) o fattori sociali.

I due diversi metodi per la determinazione del sesso, GSD ed ESD, non si escludono a vicenda; in molte specie sono state osservate transizioni tra queste due strategie (Li e Gui 2018).

1.1.1. Mammiferi

Nei mammiferi prevale il sistema di determinazione del sesso eterogametico maschile XX/XY. Il gene *Sry* (Sex-determining Region Y), situato nel cromosoma Y, è stato il primo gene determinante il sesso identificato nei vertebrati (Koopman, et al. 1991) ed è sufficiente e necessario per innescare lo sviluppo maschile. Infatti mutazioni di questo gene nell'uomo o la sua delezione nel topo determinano lo sviluppo del fenotipo sessuale femminile in individui XY; viceversa la presenza di *Sry* in topi transgenici XX porta allo sviluppo dei testicoli e alla completa inversione di sesso. *Sry* codifica per il fattore di determinazione del testicolo TDF (Testis-determining factor), una proteina che agisce come fattore di trascrizione e determina il differenziamento della gonade in senso maschile durante lo sviluppo embrionale. Questa proteina contiene un dominio di legame con il DNA caratteristico delle proteine High Mobility Group (HMG) e perciò chiamato HMG-box, una sequenza di localizzazione nucleare responsabile dell'attività di TDF; questa regione è transitoriamente espressa e innesca la differenziazione del testicolo e lo sviluppo maschile tramite differenziamento delle cellule del Sertoli. Si pensa che la proteina si leghi alle regioni regolatorie di altri geni che intervengono nel

differenziamento delle gonadi a valle di *Sry*, e nei confronti dei quali *Sry* svolgerebbe dunque la funzione di regolatore trascrizionale.

La famiglia dei geni *Sox* (Sry-related HMG box-containing genes), che codifica per un gruppo di fattori trascrizionali caratterizzati dal dominio HMG, è ampiamente coinvolta nella determinazione e differenziazione del sesso nei vertebrati. Si ipotizza che il gene *Sry* sia la diversificazione allelica del gene *Sox3*, visto che il gruppo HMG di *Sry* e *Sox3* hanno un alto livello di omologia (Sato, et al. 2010). Sebbene il gene *Sox3* non abbia una funzione primaria di determinazione del sesso nei mammiferi, l'espressione ectopica di *Sox3* durante lo sviluppo gonadico nei topi XX porta alla completa inversione del sesso da femmine XX a maschi (Sutton, et al. 2011).

Nonostante la maggior parte dei mammiferi presenti un sistema eterogametico maschile XX/XY con *Sry* come principale gene che determina il sesso, in diverse specie si sono sviluppate varianti di questo sistema. In alcuni roditori (*Tokudaia osimensis osimensis* e *Tokudaia osimensis spp.*) il cromosoma Y e il gene *Sry* sono stati persi e individui di sesso maschile e femminile hanno lo stesso cariotipo XO (Sutou, Mitsui e Tsuchiya 2001). Sebbene il percorso di differenziazione maschile risulti conservato nelle specie di mammiferi XO, ancora non si sa quali siano i geni che determinano il sesso in questi animali privi del gene *Sry*.

I sistemi di determinazione del sesso sono diversi anche in alcune specie di monotremi. Negli ornitorinchi (*Ornithorhynchus anatinus*), le femmine hanno cinque paia di cromosomi X ($X_1X_2X_3X_4X_5$), e i maschi hanno cinque cromosomi X ($X_1X_2X_3X_4X_5$) e cinque cromosomi Y ($Y_1Y_2Y_3Y_4Y_5$) (Rens, Grützner, et al. 2004). L'echidna dal becco corto (*Tachyglossus aculeatus*) ha un sistema di cromosomi sessuali simile con cinque cromosomi X ($X_1X_2X_3X_4X_5$) e quattro Y ($Y_1Y_2Y_3Y_4$) (Rens, CM O'Brien, et al. 2007). Tuttavia, questi due mammiferi monotremi non hanno il gene *Sry*, e il gene che più probabilmente determina il sesso è il gene *Amh* (anti-Müllerian hormone) sul cromosoma Y_5 (Cortez, et al. 2014). *Amh*, membro della famiglia TGF-β (Tranforming Growth Factor β), è un gene espresso dalle cellule del Sertoli che codifica l'ormone antimülleriano responsabile della regressione dei dotti di Müller negli embrioni di sesso maschile. Il differenziamento del fenotipo sessuale maschile richiede, oltre alla formazione del testicolo, la formazione dei gonodotti maschili a partire dai dotti di Wolff e la contemporanea degenerazione dei precursori dei gonodotti femminili, i dotti di Müller.

1.1.2. Uccelli

Gli uccelli presentano un sistema eterogametico femminile ZZ/ZW; il loro sesso è determinato dalla femmina e non dal maschio. Le femmine portano coppie ZW, quindi

se i figli non ereditano il cromosoma Z dalla madre, allora saranno femmine. Il sesso degli uccelli viene determinato tramite il dosaggio del gene *Dmrt1* (DM related transcription factor 1) presente sul cromosoma Z (Smith, et al. 2009). Il gene *Dmrt1* codifica per un fattore di trascrizione con un dominio di legame al DNA conservato (dominio DM) che è altamente espresso nelle gonadi maschili di pesci, rettili, uccelli e mammiferi. Il knockdown dell'espressione di *Dmrt1* tramite RNA interference (RNAi) negli embrioni di pollo porta alla femminilizzazione delle gonadi negli individui con genotipo maschile (ZZ), mentre la sovraespressione di *Dmrt1* induce il percorso maschile e antagonizza il percorso femminile delle gonadi embrionali (Li e Gui 2018).

1.1.3. Rettili

Anche nei rettili e negli anfibi è stato osservato un sistema di determinazione del sesso XX/XY; fino ad ora, oltre al gene *Sry*, non è stato individuato nessun gene per la differenziazione del sesso nei cromosomi sessuali XY.

Per i serpenti la determinazione del sesso segue il sistema ZZ/ZW. Il cromosoma Z nel serpente è omologo a un autosoma del pollo e questo indica che i cromosomi sessuali nei serpenti e negli uccelli derivano da diverse coppie autosomiche di un antenato comune.

Nei rettili risulta frequente la determinazione ambientale del sesso. La forma più comune di ESD è la determinazione del sesso dipendente dalla temperatura (TSD), ossia la temperatura ambientale durante lo sviluppo determina il sesso della prole. La TSD è stata identificata da tempo nell'agama comune (*Agama agama*), ma il meccanismo molecolare che spiega come la temperatura determina il sesso è stato svelato solo recentemente nella tartaruga dalle orecchie rosse (*Trachemys scripta*). In questo animale, il gene *Dmrt1*, essenziale per la determinazione del sesso, mostra un'espressione dipendente dalla temperatura. Il gene *Kdm6b*, regolato in modo epigenetico, svolge un ruolo determinante eliminando la trimetilazione di H3K27 (lisina, il ventisettesimo aminoacido nell'istone H3) nel promotore di *Dmrt1* (Ge, et al. 2018).

Altri fattori ambientali che regolano la determinazione del sesso possono essere: fotoperiodo, fattori sociali, ormoni esterni e sostanze nutritive (Capel 2017). Nei rettili, inoltre, si verificano spesso transizioni tra i sistemi di determinazione del sesso, specialmente nei gechi (Li e Gui 2018).

1.1.4. Anfibi

In alcune specie di anfibi, oltre al sistema di determinazione del sesso XX/XY, è stato identificato anche il sistema ZZ/ZW e varianti di questo sistema. Nella rana

neozelandese (*Leiopelma hochstetteri*), per esempio, il cromosoma Z è stato perso e il sesso è determinato tramite il cromosoma W univalente nella femmina; si hanno quindi femmine OW e maschi OO. La rana rugosa giapponese (*Rana rugosa*) mostra un sistema eterogametico sia maschile che femminile; le popolazioni settentrionali hanno il sistema ZZ/ZW e quelle meridionali hanno il sistema XX/XY. Nella rana artigliata (*Xenopus tropicalis*), invece, è stata identificata la presenza simultanea di tre diversi cromosomi sessuali, W, Z e Y; sono presenti tre tipi di maschi (YZ, YW e ZZ) e due tipi di femmine (ZW e WW).

Negli anfibi, le specie studiate fino ad ora mostrano tutte GSD (sistemi XX/XY e ZZ/ZW), ma è stata rilevata l'inversione del sesso indotta da fattori ambientali sia nelle popolazioni selvatiche che nei ceppi di laboratorio di rana comune (*Rana temporaria*) (Li e Gui 2018).

1.1.5. Pesci

I meccanismi di determinazione del sesso nei pesci sono diversi e complessi; solo poche specie presentano un sistema eterogametico (Chen, Zhu e Hu 2022).

In alcune specie ittiche è stato identificato un sistema ZZ/ZW e varianti di questo sistema. Sono stati individuati anche i corrispondenti geni determinanti il sesso. Nei pesci appartenenti alla specie *Cynoglossus semilaevis*, il gene *Dmrt1* sul cromosoma Z è stato identificato come gene che determina il sesso.

In diverse specie di pesci sono stati individuati sistemi che presentano cromosomi sessuali multipli. Nel pesce gatto (*Ancistrus sp.2*) le femmine sono $Z_1Z_2W_1W_2$ e i maschi $Z_1Z_1Z_2Z_2$; nel pesce lucertola (*Trachinocephalus myops*) sono stati scoperti sistemi in cui gli individui di sesso femminile sono ZW_1W_2 e quelli di sesso maschile ZZ (Li e Gui 2018).

Per le specie ittiche che presentano un sistema XX/XY esistono diversi geni che determinano il sesso. Questi possono essere divisi in due categorie principali.

Il primo gruppo comprende i fattori di trascrizione, per esempio *Dmrt1*, *Sox2*, *Sox3*.

I geni contenenti i domini Doublesex e Mab-3 (DM) sono coinvolti nella determinazione del sesso e nello sviluppo sessuale nei pesci, oltre che in mammiferi, uccelli, rettili, anfibi, mosche, vermi e coralli. Nel pesce medaka (*Oryzias latipes*), il gene *Dmy* (DM-domain gene on the Y chromosome) è un duplicato di *Dmrt1* sul cromosoma Y e rappresenta il principale gene determinante il sesso.

Anche la famiglia dei geni *Sox* è coinvolta nella determinazione e differenziazione del sesso nei pesci. L'ortologo del gene *Sox3* sul cromosoma Y è coinvolto nella

determinazione del sesso maschile nel pesce denominato Indian ricefish (*Oryzias dancena*). Recentemente è stato rilevato che il gene *Sox5* è coinvolto nella determinazione del sesso anche in medaka; questo gene regola negativamente l'attività di *Dmy* tramite il legame al suo promotore e mutanti di *Sox5* provocano l'inversione del sesso da femmina a maschio (Schartl, et al. 2018).

Alla seconda categoria appartengono invece i membri della via di segnale di TGF-β, che comprendono *Amhy*, *Gsdfy*, *Gdf6b*, *Amhr2*, *Bmpr1bb* e altri.

Nel pesce perjerrey (*Odontesthes hatcheri*), *Amhy*, la copia di *Amh* sul cromosoma Y, è un gene che determina il sesso maschile. Nella tilapia del Nilo (*Oreochromis niloticus*), un duplicato Y-specifico di *Amh* svolge un ruolo essenziale per la determinazione del sesso maschile. *Gsdf* (gonadal soma-derived factor), un gene della famiglia TGF-β, è stato identificato nei pesci e si presume che svolga un ruolo importante nello sviluppo del testicolo. Nel pesce Luzon ricefish (*Oryzias luzonensis*), *Gsdfy* ha sostituito *Dmy* come principale gene determinante del sesso. *Gdf6y* (growth differentiation factor 6) è espresso transitoriamente poco dopo la schiusa ed è stato proposto come principale gene che determina il sesso nel killifish (*Nothobranchius furzeri*).

Rispetto ai sistemi di determinazione genetica del sesso (GSD) relativamente stabili nei mammiferi e negli uccelli, i sistemi di determinazione del sesso negli ectotermi mostrano grande diversità; questi sistemi comprendono GSD eterogametica maschile, GSD eterogametica femminile e determinazione ambientale del sesso (ESD). Nei pesci, come in rettili e anfibi, si verificano frequentemente transizioni tra diverse strategie di determinazione del sesso (Capel 2017). Alcune specie ittiche possono cambiare sesso tramite interazione con individui appartenenti alla stessa specie. Nel pesce gobide (*Trimma okinawae*), la più grande inversione del sesso da femmina a maschio avviene dopo la rimozione del maschio dominante da un gruppo riproduttivo, ma i maschi neo-formati conservano sia le ovaie che i testicoli e possono cambiare sesso e tornare femmine in presenza di maschi più grandi nel gruppo (Li e Gui 2018).

1.2. Mugil cephalus

Il cefalo (*Mugil cephalus*) è la specie più diffusa della famiglia Mugilidae, che comprende un totale di 20 generi e 70 specie.

Tabella 1. *Classificazione scientifica del cefalo.*

Classificazione scientifica	
Dominio	Eukarya
Regno	Animalia
Sottoregno	Eumetazoa
Superphylum	Deuterostomia
Phylum	Chordata
Subphylum	Vertebrata
Infraphylum	Gnathostomata
Superclasse	Osteichthyes
Classe	Actinopterygii
Sottoclasse	Teleostei
Superordine	Acanthopterygii
Ordine	Mugiliformes
Famiglia	Mugilidae
Genere	Mugil
Specie	M. cephalus

1.2.1. Caratteristiche morfologiche

Il cefalo è dotato di un corpo robusto ed allungato, possiede una testa larga, appiattita centralmente ed una bocca piccola con un labbro superiore sottile e liscio. L'osso mascellare è dritto e la sua estremità posteriore non risulta visibile quando il pesce mantiene la bocca chiusa. Gli occhi sono ricoperti da una palpebra adiposa che si estende anteriormente e posteriormente ad esso, lasciando libera soltanto una piccola fessura centrale. Le due pinne dorsali sono corte e le pettorali sono inserite in una posizione piuttosto alta. Il corpo del cefalo mostra tonalità grigio bluastre nella zona dorsale, una colorazione argentea con linee longitudinali grigie sui fianchi ed una pigmentazione più chiara, tendente al bianco argenteo nella porzione ventrale. Alla base delle pinne pettorali è presente una macchia scura (https://www.agraria.org/pesci/muggine.htm).

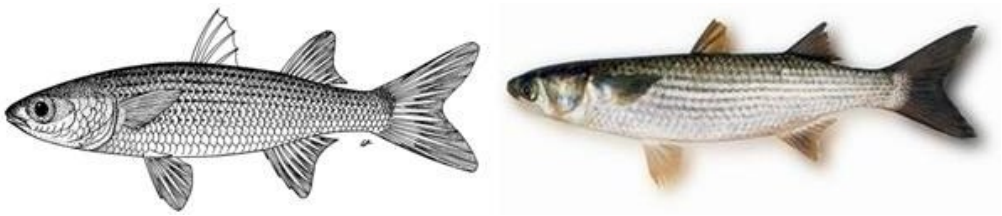


Figura 1. *Mugil cephalus*

1.2.2. Habitat

M. cephalus è una specie cosmopolita che vive nelle acque costiere dei principali oceani delle zone tropicali, subtropicali e temperate, principalmente tra le latitudini 42 N e 42 S. Nelle aree occidentali dell'Oceano Atlantico, il cefalo risulta molto presente, dalle acque della Nuova Scozia (Canada) fino a quelle del Brasile, compreso il Golfo del Messico. Nelle zone orientali dell'Atlantico, è presente dalle acque francesi fino a quelle del Sud Africa ed è comune anche nel Mar Mediterraneo e nel Mar Nero (https://www.agraria.org/pesci/muggine.htm).

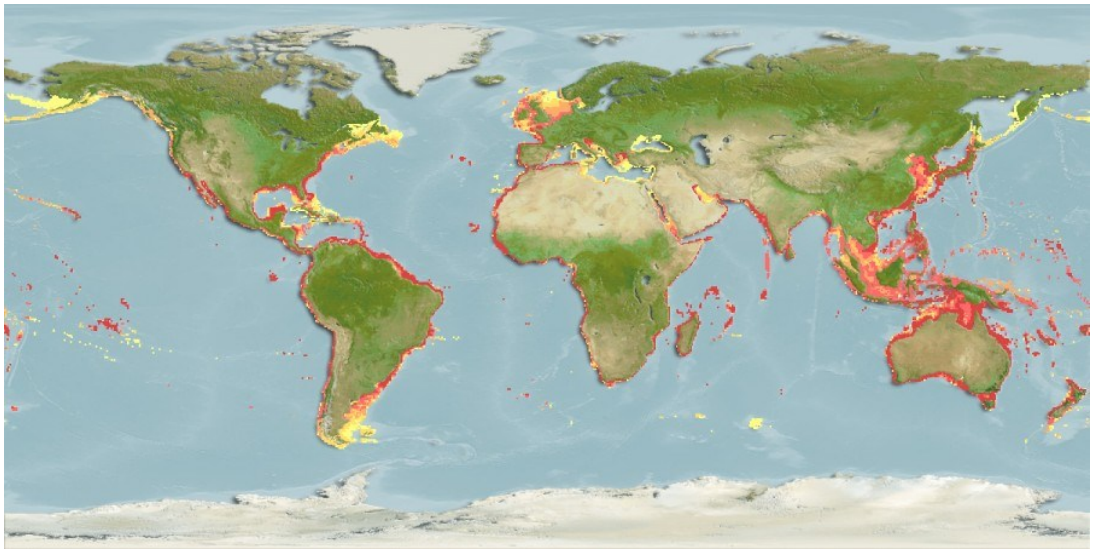


Figura 2. *Habitat di M. cephalus. Le zone contrassegnate con il colore rosso rappresentano i principali habitat del cefalo.*

Questo pesce è caratterizzato da comportamento catadromo (discende le correnti dei fiumi per riprodursi e deporre le uova in mare) e la sua presenza viene abitualmente riscontrata lungo la costa negli estuari e negli ambienti d'acqua dolce. Cefali adulti sono stati trovati in acque che vanno dallo 0 al 75 % di salinità, mentre i giovani possono tollerare intervalli di salinità così ampi solo dopo aver raggiunto una lunghezza di 4-7 cm (https://www.fao.org/).

1.2.3. Riproduzione

Gli adulti, durante la stagione riproduttiva (mesi autunnali e invernali), formano enormi banchi e migrano verso il mare aperto per deporre le uova. Le femmine di cefalo possono deporre da 0,5 a 2,0 milioni di uova, a seconda delle dimensioni corporee. La schiusa avviene circa 48 ore dopo la fecondazione, liberando larve lunghe circa 2,4 mm. In seguito alla schiusa delle uova, le larve si spostano verso la costa in acque estremamente basse, all'interno di siti costieri riparati, dove trovano abbondanti quantitativi di cibo e protezione dagli attacchi di pesci predatori. Dopo aver raggiunto la lunghezza di circa 5 cm, si spostano in acque leggermente più profonde (https://www.fao.org/).

1.2.4. Alimentazione

Il cefalo si nutre durante il giorno e consuma principalmente zooplancton, piccole larve di insetti e materiale vegetale in decomposizione. Inoltre, grazie alla conformazione dello stomaco e al suo lungo tratto gastrointestinale, è molto abile nel digerire detriti di vario genere. Per poter triturare al meglio questi materiali il cefalo ingerisce vari tipi di sedimenti, che all'interno del suo particolare stomaco, vengono utilizzati per sminuzzare i detriti. Secondo alcune ricerche, il quantitativo di sedimento riscontrato nello stomaco di questo mugilide aumenta in relazione alla grandezza degli animali, dimostrando il fatto che l'abitudine detritivora risulta accentuata soprattutto negli esemplari adulti. La forma larvale si nutre principalmente di microcrostacei (https://www.agraria.org/pesci/muggine.htm).

1.2.5. Allevamento e interesse economico

La maggior parte degli avannotti viene raccolta allo stato selvatico, in particolare nel Mediterraneo orientale e meridionale, in Arabia Saudita e negli Stati del Golfo e nel sud-est asiatico. L'Egitto è il più grande produttore di cefalo d'allevamento e la produzione è aumentata rapidamente durante il periodo 1998-2003. Anche la Repubblica di Corea, l'Italia, la provincia cinese di Taiwan e Israele sono produttori importanti.

I cefali vengono catturati dalle zone costiere di basso fondale e dagli estuari dei fiumi nel periodo che va da maggio alla prima metà di dicembre, per essere successivamente introdotti in allevamento.

M. cephalus generalmente viene immesso sul mercato in seguito a processi di salatura ed affumicatura e la bottarga salata, tra le varie forme di commercializzazione, rappresenta il prodotto che viene maggiormente apprezzato (https://www.agraria.org/pesci/muggine.htm). La domanda di uova di cefalo è cresciuta

notevolmente in molte parti del mondo negli ultimi decenni, tanto che il cefalo viene anche chiamato “oro grigio” dai pescatori (Hung e Shaw 2006).

1.2.6. La bottarga

La bottarga di muggine è un prodotto ottenuto da salagione ed essiccazione delle ovaie di *M. cephalus*. Il colore va dall’ambrato chiaro all’ambrato scuro. Viene commercializzata intera o macinata e il suo prezzo è compreso tra i 150 e i 300 €/Kg.

La tecnologia produttiva prevede le seguenti fasi:

- Cattura, selezione ed eviscerazione del pesce
  Dopo la cattura, il cefalo viene separato dalle altre specie, e successivamente vengono divisi gli individui di sesso maschile da quelli di sesso femminile. Le ovaie vengono prelevate subito dopo l’eviscerazione del pesce.
- Lavaggio e preparazione delle ovaie alla successiva salagione
  L’eviscerazione è immediatamente seguita dalla toelettatura dell’ovario e dallo svuotamento manuale delle vene ovariche al fine di prevenire la formazione di un “pabulum” ideale per lo sviluppo microbico e la comparsa di alterazioni nel prodotto finito.
- Salagione delle ovaie
  Dopo il lavaggio le ovaie vengono trattate a secco con sale marino.
- Pressatura delle ovaie
  Al termine della salagione le ovaie, lavate in soluzione satura salina, vengono disposte in strati su ripiani in legno e pressate per favorire l’allontanamento della salamoia e dei liquidi in eccesso (“spurgo”) e consentire la penetrazione e la diffusione uniforme del sale.
- Stagionatura
  Le ovaie dopo la pressatura sono trasferite nella sala di stagionatura dove vengono essiccate fino ad ottenere la perdita di peso desiderata (30-50% del peso). A fine stagionatura la bottarga presenta colore uniforme, membrana perfettamente aderente all’ovario e difficilmente asportabile.
- Dopo la stagionatura e prima dei essere confezionate, le ovaie vengono lavate superficialmente per eliminare eventuali residui di sale e fatte asciugare.
- Confezionamento
  Il prodotto finale viene leggermente oliato, confezionato sottovuoto o grattugiato in barattoli.

La tecnica di preparazione della bottarga è nota sin dall'antichità spesso come attività secondaria dei pescatori di tonno e di muggine. Sembra certo che siano stati i fenici e poi gli arabi a diffondere la produzione ed il consumo di questo prodotto in tutta l'area mediterranea; lo stesso termine "bottarga" sembra derivare dall'arabo "Butarih" che significa pesce salato o affumicato.

Le uova di muggine salate a secco sono probabilmente il prodotto a base di uova essiccate più conosciuto. Nei diversi paesi questo alimento è conosciuto con nomi locali, ad esempio Avgotaraho in Grecia, Poutargue in Francia, Bottarga di Muggine in Italia, Batarekh in Egitto e Libano, Yu chian-cha in Cina, Wuoze in Taiwan e Karasumi in Giappone. La Sardegna è la principale regione italiana per la produzione di bottarga di muggine (Usman, et al. 2022).

Figura 3. *Bottarga di muggine.*

1.3. La determinazione del sesso in *Mugil cephalus*

I pesci teleostei presentano una grande varietà di meccanismi di determinazione del sesso che comprendono fattori genetici, ambientali e sociali (L. Dor, et al. 2016). Quando si osserva la determinazione genetica del sesso (GSD), i meccanismi variano anche tra specie strettamente imparentate o all'interno della stessa specie, e fattori ambientali possono influenzare la determinazione genetica del sesso (Ferraresso, et al. 2021).

L'alto valore commerciale delle uova di muggine ha portato ad un incremento dell'interesse per i fenomeni alla base della determinazione del sesso del cefalo con lo scopo di ottenere popolazioni di sole femmine (L. Dor, et al. 2016). L'obiettivo è quello di sviluppare allevamenti intensivi di cefalo per la produzione di uova, come già realizzato a Taiwan (Crosetti 2016). L'ideale sarebbe trovare un metodo rapido e non letale per lo smistamento precoce di maschi e femmine, anche prima della maturazione sessuale,

così da poter allevare ceppi monosessuali di femmine per poter utilizzare le loro uova (Ferraresso, et al. 2021).

Nel corso degli ultimi anni sono stati condotti diversi studi con lo scopo di individuare possibili geni che permettano di effettuare una precoce determinazione del sesso.

In uno studio condotto da Dor et al. nel 2016 è stato usato un approccio basato sulla sintenia per selezionare marcatori microsatelliti potenzialmente dispersi nel genoma del cefalo basandosi sul genoma della tilapia del Nilo (*O. niloticus*). Lo studio ha permesso di individuare cinque geni potenzialmente associati alla determinazione del sesso: *gth-ri*, *foxi1*, *dhx32*, *bub3* e *dock1*. Il gene *gth-ri* (recettore della gonadotropina I), anche conosciuto come *fshr* (che codifica per i recettori dell'ormone follicolo-stimolante), è noto per avere un ruolo chiave nella follicologenesi nelle gonadi femminili di pesci e mammiferi (Li e Cheng 2018). La sua espressione è stata trovata nelle gonadi dei pesci e localizzato nelle cellule della teca e della granulosa. Nello stesso studio è stato inoltre evidenziato che il sesso della prole è determinato solo dagli alleli trasmessi dal padre e questo indica un sistema di determinazione del sesso XY (L. Dor, et al. 2016).

Lo stesso gruppo di ricerca, nel 2020 ha individuato *bccip*, *dhx32a*, *dock1* e *fshr* (*gth-ri*) come geni candidati per essere regolatori della determinazione del sesso (master key regulators of SD) (L. Dor, et al. 2020).

Nello studio condotto da Ferraresso et al. (2021), è stato implementato un approccio di sequenziamento dell'intero genoma (whole-genome sequencing) attraverso Pool Sequencing (Pool-Seq) con l'obiettivo di identificare marcatori genetici associati al sesso, che sono stati poi validati su quattro popolazioni selvatiche. Gli individui che sono stati analizzati erano cefali (*M. cephalus*) adulti maschi e femmine appartenenti a 4 diverse popolazioni provenienti dall'area mediterranea: Cabras (CAB) in Sardegna (ovest), Tortolì (TOR) in Sardegna (est), Orbetello (ORB) in Toscana e Baia di Kavala (KAV) in Macedonia.

Figura 4. *Area di provenienza degli individui analizzati.*

Per l'analisi Pool-seq sono stati utilizzati un pool di 60 femmine e un altro pool di 60 maschi. Ciascun pool comprendeva 30 individui per ciascuna delle due popolazioni sarde (TOR e CAB).

Tra le varianti che determinano il sesso, sono stati identificati 3 SNP nelle posizioni 179 (MuCe179), 206 (MuCe206) e 322 (MuCe322) del contig_111122. Le mutazioni MuCe179 e MuCe206 rappresentano varianti missenso, la mutazione MuCe322 è una variante sinonima. Le tre varianti si trovano sull'esone 14 del gene *fshr*, supportando quindi l'ipotesi che questo gene sia probabilmente coinvolto nella GSD nel cefalo.

Per convalidare i risultati del Pool-Seq, una regione di 308 nucleotidi comprendente le varianti MuCe179, MuCe206 e MuCe322 è stata sequenziata in 245 individui raccolti in quattro diversi siti che rappresentano il Mediterraneo occidentale (Laguna di Cabras), il Tirreno (Tortolì e lagune di Orbetello) e il Mar Egeo Settentrionale (Kavala). Gli individui inclusi nelle popolazioni TOR e CAB sono stati utilizzati sia per l'analisi Pool-Seq che per la convalida. L'allele *fshr* più frequente è stato considerato wild-type (wt), l'allele con tutte e tre le varianti (MuCe179, MuCe206, MuCe322) è stato denominato m1, l'allele contenente solo le due mutazioni missenso (MuCe179 e MuCe206) è stato chiamato m2, l'allele con la sola mutazione missenso MuCe179 ha preso il nome di m3 e, infine, l'allele con la sola mutazione sinonima MuCe322 è m4. Complessivamente, i dati sul genotipo *fshr* hanno confermato che gli alleli m1 e m2 sono significativamente associati al sesso maschile. Tuttavia, l'associazione non risultava completa, poiché il fenotipo sessuale e il genotipo *fshr* non corrispondevano al modello previsto nel 3-13% delle femmine.

Dallo studio si conclude che il genotipo wt/wt è significativamente associato al sesso femminile e le due varianti missenso nel locus *fshr* sono associate al sesso maschile. Il

modello osservato di maschi eterozigoti wt/m1 e wt/m2 suggerisce un tipo di GSD "XX-XY". I due SNP, presenti in eterozigosi (wt/m1 o wt/m2) nel fenotipo maschile, hanno mostrato penetranza incompleta in un certo numero di maschi che esibiscono un genotipo wt/wt, associato invece al fenotipo femminile. Questi individui sono stati denominati maschi “non conformi”.

La differenza tra le popolazioni nella frequenza delle femmine wt/m1 non è significativa ($p > 0,5$). I maschi omozigoti wild-type (wt/wt) (“non conformi”) sono invece presenti a frequenze variabili nelle quattro popolazioni (6%–45%). Le differenze tra le popolazioni sono altamente significative, con la frequenza più bassa (6%) in ORB e la più alta in KAV (45%).

È stato ipotizzato che le varianti osservate nel locus *fshr* possano agire indirettamente sulla determinazione del sesso, aumentando la sensibilità della via di segnalazione che determina il sesso a fattori ambientali, ad esempio alla temperatura dell'acqua. Varianti genetiche associate al sesso influenzate dalla temperatura sono già state segnalate nella tilapia del Nilo (Wessels, et al. 2014) ed in zebrafish (Ribas , et al. 2017).

Inoltre la distribuzione divergente dei genotipi *fshr* nei maschi tra KAV (Nord Egeo) e TOR-CAB-ORB (Mediterraneo occidentale) potrebbe essere influenzata da fattori ambientali. È possibile che nella popolazione KAV il ruolo di *fshr* sia meno rilevante e che altri loci, legati a condizioni locali, contribuiscano alla determinazione del sesso (Ferraresso, et al. 2021).

In uno studio successivo, svolto dallo stesso gruppo di ricerca, sono stati riportati i risultati ottenuti dal sequenziamento di singoli individui, che ha permesso di assegnare in modo preciso il genotipo di ciascun animale, obiettivo che con Pool-seq non era stato possibile raggiungere. Nello studio sono stati sequenziati solamente individui di sesso maschile, 20 maschi non conformi (genotipo wt/wt) e 20 maschi conformi (genotipo wt/m1).

In questo lavoro sono state confermate le varianti già riscontrate precedentemente (Ferraresso, et al. 2021), ma queste sono state denominate in maniera differente in quanto è stata utilizzata una versione aggiornata del genoma di *M. cephalus* (MuCe179, MuCe206 e MuCe322 corrispondono alle varianti 385101, 385128 e 385244). Oltre a queste, l’analisi della regione FSHR±5kb ha permesso di identificare 3 nuovi SNP nelle posizioni 385481, 385494 e 387492. I primi due SNP si trovano nel gene *fshr* e risultano essere presenti nel 95% dei maschi conformi analizzati e nel 16% dei maschi non conformi. Lo SNP in posizione 387492 si trova invece nell’immediata regione a valle e

mostra una differente frequenza allelica tra i maschi non conformi di Tirreno ed Egeo. I maschi non conformi dell'Egeo, infatti, presentano la mutazione in omozigosi nel 100% dei casi, mentre nei maschi non conformi del Tirreno la mutazione è presente in omozigosi nel 50% dei casi e il restante 50% presenta la mutazione in eterozigosi come le femmine, riducendo quindi l'utilità della variante in un'ottica di differenziazione tra i sessi.

1.4. Sequenziamento

Il sequenziamento del DNA è il processo mirato a determinare la sequenza di un dato frammento di DNA, ovvero l'ordine dei nucleotidi che lo compongono.

Nel corso degli anni si sono susseguite diverse strategie per determinare la sequenza nucleotidica di un frammento di DNA.

Nel 1975 Sanger e Coulson hanno sviluppato il metodo 'plus and minus' per il sequenziamento del DNA. Due anni dopo, usando questo approccio Sanger ha determinato il genoma di 5368 pb del fago φX174 e questo è stato il primo genoma ad essere sequenziato. Lo stesso anno, ha sviluppato un nuovo metodo che può decodificare frammenti di circa quattrocento basi in un giorno (Giani, et al. 2020).

Il metodo "Sanger", chiamato anche metodo della terminazione di catena è un metodo enzimatico e si basa sull'utilizzo dei nucleotidi dideossitrifosfato (ddNTPs), molecole artificiali corrispondenti ai 4 nucleotidi naturali, ma che si differenziano da esse per l'assenza del gruppo idrossilico sul carbonio 3'. I ddNTPs, a causa della loro struttura, impediscono che un altro nucleotide si leghi ad essi; quindi l'aggiunta di un ddNTP sul filamento del DNA in estensione ne causa la terminazione prima del raggiungimento della fine della sequenza del DNA stampo. Ciò dà origine ad una serie di frammenti di DNA di lunghezza diversa interrotti in corrispondenza dell'incorporazione del ddNTP, che viene marcato con ^{32}P. I frammenti generati vengono poi fatti correre su gel per la lettura della sequenza (Sanger, Nicklen e Coulson 1977).

Il metodo "Sanger" è stato utilizzato per sequenziare genomi di fagi e virus, fino ad arrivare al sequenziamento dell'intero genoma umano, grazie al Progetto Genoma Umano (HGP), concluso nel 2003 (Giani, et al. 2020).

1.4.1. Sequenziamento di seconda generazione (Next-Generation Sequencing, NGS)

All'inizio degli anni 2000 sono state introdotte le tecnologie di seconda generazione (Next-Generation Sequencing, NGS).

Tutti gli approcci NGS si basano sulla preparazione di una libreria genomica utilizzando DNA nativo o amplificato. Dopo la frammentazione del DNA e la selezione della dimensione del frammento, vengono legati degli adattatori alle estremità di ciascun frammento; infine si procede con la fase di amplificazione del DNA. La libreria risultante viene poi caricata su una cella a flusso e sequenziata. La reazione di sequenziamento solitamente comporta l'incorporazione di deossinucleotidi marcati in una catena di DNA immobilizzata su una superficie.

I nuovi metodi di sequenziamento hanno diversi vantaggi rispetto ai precedenti, tra i quali il fatto di essere molto più economici e rapidi (Giani, et al. 2020). A differenza del metodo Sanger, le tecnologie di sequenziamento di seconda generazione, producono molti più dati; possono inoltre generare più copie delle stesse posizioni genomiche e per questo vengono anche chiamate tecnologie di sequenziamento "profonde". Le principali tipologie di piattaforme NGS prodotte nel corso degli anni sono: Roche 454, Ion Torrent, SOLiD e Illumina.

I metodi di sequenziamento di seconda generazione hanno permesso di superare diversi limiti appartenenti alle metodologie precedenti, ma presentano comunque alcuni svantaggi: la preparazione dei campioni risulta costosa in termini di tempo e di denaro, c'è la necessità di eseguire un passaggio di amplificazione che può portare ad errori, e i sistemi per rilevare i nucleotidi incorporati sono costosi. Ad oggi, tra queste metodologie, l'unica ancora rilevante per il suo utilizzo è la tecnologia Illumina.

1.4.2. Sequenziamento di terza generazione

Per superare i problemi delle tecniche NGS sono stati sviluppati nuovi metodi di sequenziamento, che prendono il nome di sequenziamento di terza generazione. Questi si differenziano dai precedenti principalmente per l'assenza di frammentazione del DNA, per il sequenziamento diretto, ossia senza pre-amplificazione e per la possibilità di sequenziare molecole più lunghe.

Le tecniche di sequenziamento di terza generazione includono la tecnologia SMRT sviluppata da PacBio e la tecnologia Oxford nanopore (ONT) basata su metodi nanopore.

Il sequenziamento tramite PacBio, definito anche SMRT (single-molecule real-time sequencing), si avvale dell'utilizzo di una DNA polimerasi ad alta processività, capace di aggiungere 1000 pb/s (paia di basi al secondo). La sequenza nucleotidica è determinata in tempo reale rilevando la reazione che si verifica in ogni base mediante un cambiamento della fluorescenza.

La tecnologia Oxford nanopore (ONT), invece, è un metodo di sequenziamento che non utilizza la fluorescenza, ma legge il segnale di ogni base mediante corrente elettrica.

Sebbene le tecnologie di terza generazione riescano a superare alcuni problemi riscontrati nella seconda generazione, ad oggi il principale fattore limitante è un alto tasso di errori e di conseguenza una sequenza non accurata. Tuttavia, queste tecnologie e gli strumenti bioinformatici ad esse correlate sono attualmente sottoposte ad un continuo perfezionamento per aumentarne la precisione ed i risultati sono promettenti per il futuro (Kang, Kang e Kim 2019).

1.4.3. Sequenziamento Illumina

La tecnologia Illumina è caratterizzata da un sequenziamento per sintesi (SBS) basato su Bridge Amplification ed è, ad oggi, la tecnologia predominante tra le piattaforme di seconda generazione. Questa tecnologia permette di produrre una grande quantità di dati con prezzi di esecuzione bassi ed è quindi ampiamente utilizzata nel sequenziamento di grandi volumi di DNA. La piattaforma Illumina ha un basso tasso di errore con una precisione complessiva superiore al 99,5%; presenta però una tendenza a commettere errori di sostituzione nel sequenziamento di regioni ricche di AT e GC (Kang, Kang e Kim 2019).

Il primo passaggio necessario per eseguire il sequenziamento consiste nella preparazione delle librerie, durante la quale il DNA viene frammentato in segmenti di una data lunghezza (solitamente compresi nel range 200-500 nt) a cui verranno aggiunti adattatori specifici ad entrambe le estremità.

Il costrutto finale contiene:

- gli adattatori (sequenze p5 e p7), ossia sequenze complementari agli oligonucleotidi presenti sulla superficie della flowcell nella quale avvengono l'amplificazione e il sequenziamento; questi consentono quindi alla libreria di legarsi e generare cluster sulla flowcell
- le sequenze dei siti di legame dei primer per avviare il sequenziamento (Read1 SP e Read2 SP)
- le sequenze degli identificatori univoci di ciascun campione (Index 1 e Index 2) che consentono di multiplexare più campioni in una stessa reazione di sequenziamento.

Figura 5. *Rappresentazione schematica del costrutto.*

In seguito viene eseguita un'amplificazione delle sequenze, in modo da fornire un segnale sufficiente per il sequenziamento.

Il passaggio successivo prevede la generazione di cluster. I frammenti vengono cioè clonati in migliaia di copie identiche. Per generare i cluster, la libreria viene caricata in una flowcell, cioè un supporto solido con diverse corsie la cui superficie è ricoperta da oligonucleotidi complementari agli adattatori presenti sulle librerie (p5 e p7). L'ibridizzazione è permessa dagli oligonucleotidi presenti nella superficie.

Una volta legata alla flowcell, la libreria viene estesa dalla polimerasi per ottenere cluster con migliaia di copie della molecola iniziale. Per fare ciò vengono eseguiti i seguenti passaggi:

- la molecola a doppio filamento viene denaturata e il filamento stampo originale viene lavato via
- la molecola a filamento singolo, legata per un'estremità alla flowcell, si piega e forma un ponte legandosi ad un primer complementare adiacente
- una polimerasi sintetizza il frammento reverse, formando un ponte a doppio filamento
- il ponte a doppio filamento viene denaturato, formando due copie di DNA stampo legate alla superficie della flowcell.

Questo processo viene ripetuto diverse volte e si verifica simultaneamente per milioni di clusters con conseguente amplificazione clonale di tutti i frammenti. Dopo la bridge amplification, i frammenti reverse vengono lavati via, lasciando solo i frammenti forward. Le estremità 3' libere vengono bloccate per impedire la nuova ibridazione con i primer ancorati alla superficie.

L'ultimo passaggio, ossia il sequenziamento, inizia con l'introduzione del primo primer di sequenziamento nella flowcell e la sua ibridazione all'adattatore. Ad ogni ciclo poi, vengono aggiunti i nucleotidi marcati mediante fluorofori che competono per essere addizionati alla catena in crescita. Soltanto il nucleotide complementare alla sequenza modello viene incorporato, mentre gli altri vengono lavati via. Dopo l'aggiunta di ogni nucleotide, i cluster vengono eccitati da una sorgente luminosa e, a seconda della base che viene aggiunta alla catena, viene emesso un caratteristico segnale fluorescente. La

specifica lunghezza d'onda, assieme all'intensità del segnale, permette di determinare la base presente.

Dopo il completamento della prima read, il prodotto letto viene lavato via e il terminatore viene rimosso. Il modello poi si piega nuovamente e il secondo indice viene letto alla stessa maniera del primo. Le polimerasi estendono il secondo oligonucleotide della flow cell formando un ponte a doppio filamento, che viene linearizzato e vengono bloccate le estremità 3'. Il frammento originale forward viene lavato via lasciando solo il frammento reverse. La read 2 inizia con l'introduzione dell'apposito primer e i passaggi per il sequenziamento vengono ripetuti come per la read 1.

Questo processo genera milioni di reads che rappresentano tutti i frammenti. Le sequenze provenienti da librerie composte da più campioni miscelati, vengono poi separate sulla base degli identificatori univoci (index) introdotti durante la preparazione dei campioni. Le letture forward e reverse sono successivamente appaiate per creare sequenze contigue.

Nella fase di sequenziamento, esistono 3 diverse metodologie per differenziare le basi; a seconda del sistema utilizzato, si può avere un sequenziamento con

- chimica a quattro canali: vengono utilizzati quattro coloranti fluorescenti, uno per ogni base, e quattro immagini per ciclo di sequenziamento
- chimica a due canali: vengono utilizzati due coloranti fluorescenti e due immagini per ciclo di sequenziamento per codificare i dati per le quattro basi del DNA. Sono presenti un'immagine dal canale rosso e un'immagine dal canale verde; una mancata identificazione viene indicata con una N
- chimica a un canale: utilizza la tecnologia CMOS per l'identificazione delle basi utilizzando due immagini per ciclo di sequenziamento. I nucleotidi vengono identificati tramite l'analisi dei diversi schemi di emissione per ogni base attraverso le due immagini. L'adenina presenta una marcatura rimovibile ed è marcata solo nella prima immagine, la citosina presenta un gruppo che si lega a una marcatura ed è marcata solo nella seconda immagine, la timina presenta una marcatura fluorescente ed è dunque marcata in entrambe le immagini, mentre la guanina è sempre scura (non marcata) (https://www.illumina.com/).

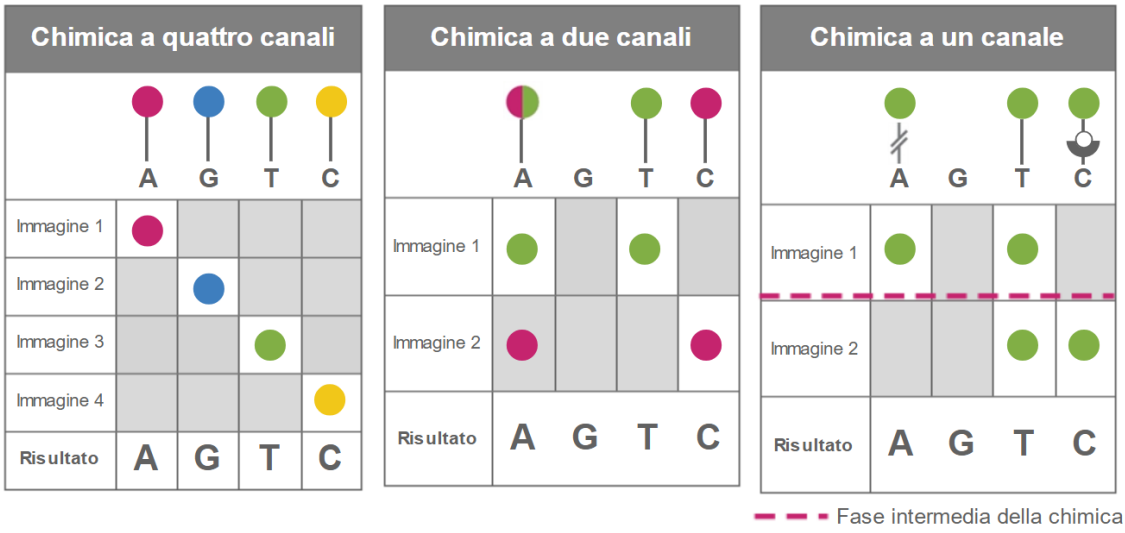


Figura 6. *Rappresentazione schematica delle tre differenti metodologie utilizzate nella fase di sequenziamento.*

Il sequenziamento con la metodologia Illumina o con le altre tecnologie NGS consente di eseguire l'analisi dell'intero genoma o Whole Genome Sequencing (WGS). Questo può servire per individuare le differenze presenti all'interno dei genomi di differenti individui appartenenti ad una determinata specie e quindi identificare possibili polimorfismi che caratterizzano ciascun individuo o gruppo di individui.

I polimorfismi che vengono individuati prendono il nome di varianti a singolo nucleotide o SNP (Single Nucleotide Polymorphism). Si tratta di mutazioni puntiformi, cioè una variazione della sequenza del DNA che interessa uno o pochi nucleotidi. La sostituzione di una base nucleotidica può avere diverse conseguenze sulla proteina codificata dal gene e in base a queste, vengono distinte le mutazioni in: silenti, quando la sostituzione nucleotidica porta alla formazione di una nuova tripletta che codifica per lo stesso amminoacido, quindi non si verifica un cambiamento della proteina prodotta; neutre, quando la tripletta mutata codifica per un amminoacido che mantiene le stesse funzioni dell'amminoacido originale; missenso, nel caso in cui la nuova tripletta codifichi per un amminoacido differente, portando alla formazione di una proteina diversa; nonsenso se il nuovo codone che si forma dalla sostituzione codifica per il segnale di stop, con conseguente formazione di una proteina tronca, quindi non funzionale.

2. Scopo del lavoro

L'interesse economico legato al cefalo (*M. cephalus*) è rivolto principalmente agli individui di sesso femminile per la produzione della bottarga, una specialità culinaria costituita dalla gonade femminile pressata, salata ed essiccata. Negli ultimi decenni la domanda di uova di cefalo è cresciuta notevolmente in molte parti del mondo. Questo ha portato alla necessità di comprendere i meccanismi alla base della determinazione del sesso per poter ottenere popolazioni di sole femmine (L. Dor, et al. 2016). Questo porterebbe significativi vantaggi economici all'industria del cefalo, andando a ridurre i costi per l'allevamento di individui di sesso maschile.

Nel corso degli ultimi anni sono stati condotti diversi studi volti ad individuare i possibili geni coinvolti nella determinazione del sesso. Dor et al. hanno indicato i geni *gth-ri* (*fshr*), *foxi1*, *dhx32*, *bub3*, *dock1*, *bccip* e *dhx32a* come possibili geni coinvolti nella determinazione del sesso (L. Dor, et al. 2016; L. Dor, et al. 2020). In un lavoro condotto da Ferraresso et al. nel 2021 sono stati identificati tre SNP potenzialmente associati alla determinazione del sesso, MuCe179, MuCe206 e MuCe322, sull'esone 14 del gene *fshr*. Le mutazioni MuCe179 e MuCe206, che rappresentano varianti missenso, risultano associate al fenotipo maschile; hanno però mostrato penetranza incompleta in un certo numero di maschi che esibiscono un genotipo wt/wt, associato invece al fenotipo femminile. Questi individui sono stati denominati maschi "non conformi" (Ferraresso, et al. 2021). La percentuale di maschi non conformi si è rivelata notevolmente variabile a seconda dell'origine geografica degli animali, passando dal 6% nel Tirreno orientale al 45% nell'Egeo. In uno studio successivo, svolto dallo stesso gruppo di ricerca, è stato eseguito il sequenziamento dell'intero genoma di individui di sesso maschile, suddivisi in conformi e non conformi, per entrambe le popolazioni. Oltre alla conferma delle varianti già riscontrate, l'analisi della regione FSHR±5kb ha permesso di identificare tre nuovi SNP nelle posizioni 385481, 385494 e 387492 in grado di aumentare l'accuratezza nella distinzione tra maschi conformi e non conformi. Il potenziale di questi ultimi SNP nella discriminazione tra sesso maschile e sesso femminile non era però ancora stata investigata su individui di sesso femminile.

L'obiettivo del presente lavoro di tesi è quello di ampliare lo studio in corso, includendo l'analisi delle femmine, allo scopo di individuare ulteriori mutazioni associate al gene *fshr* che possano contribuire alla determinazione del sesso. Sono stati quindi sequenziati i genomi di 32 femmine appartenenti ad entrambe le popolazioni. I dati di sequenziamento acquisiti sono stati analizzati assieme a quelli precedentemente ottenuti per i 40 individui di sesso maschile allo scopo di verificare la presenza di SNP coinvolti nella determinazione del sesso.

3. Materiali e metodi

3.1. Campioni utilizzati nello studio

Per il presente studio sono stati utilizzati complessivamente 72 campioni di *Mugil cephalus*, suddivisi in due popolazioni provenienti da differenti aree geografiche: Egeo (KAV) e Tirreno (TIR). Per ciascuna popolazione, sono stati analizzati 16 individui di sesso femminile e 20 di sesso maschile. I dati degli individui di sesso maschile, suddivisi in maschi conformi e maschi non conformi, provengono da un precedente lavoro fatto dal gruppo di ricerca presso cui ho svolto la mia tesi, mentre per le femmine sono stati eseguiti l'estrazione del DNA e il sequenziamento.

Nella tabella 2 vengono riportate le caratteristiche dei campioni utilizzati.

Tabella 2. *Caratteristiche dei campioni analizzati.*

Gruppo	Campione	Luogo di provenienza	Genotipo
KAV femmine	KAV_F02	Baia di Kavala	wt/m1
	KAV_F04	Baia di Kavala	wt/m1
	KAV_F06	Baia di Kavala	wt/wt
	KAV_F07	Baia di Kavala	wt/wt
	KAV_F09	Baia di Kavala	wt/wt
	KAV_F10	Baia di Kavala	wt/wt
	KAV_F11	Baia di Kavala	wt/wt
	KAV_F12	Baia di Kavala	wt/wt
	KAV_F21	Baia di Kavala	wt/wt
	KAV_F22	Baia di Kavala	wt/wt
	KAV_F24	Baia di Kavala	wt/wt
	KAV_F25	Baia di Kavala	wt/m1
	KAV_F27	Baia di Kavala	wt/m1
	KAV_F29	Baia di Kavala	wt/wt
	KAV_F30	Baia di Kavala	wt/wt
	KAV_F31	Baia di Kavala	wt/wt
KAV maschi conformi	KAV_M02	Baia di Kavala	wt/m1
	KAV_ M03	Baia di Kavala	wt/m1
	KAV_ M04	Baia di Kavala	wt/m1
	KAV_ M11	Baia di Kavala	wt/m1
	KAV_ M13	Baia di Kavala	wt/m1
	KAV_ M15	Baia di Kavala	wt/m1
	KAV_ M16	Baia di Kavala	wt/m1
	KAV_ M18	Baia di Kavala	wt/m1

	KAV_ M19	Baia di Kavala	wt/m1
	KAV_ M20	Baia di Kavala	wt/m1
KAV maschi non conformi	KAV_ M01	Baia di Kavala	wt/wt
	KAV_ M05	Baia di Kavala	wt/wt
	KAV_ M07	Baia di Kavala	wt/wt
	KAV_ M08	Baia di Kavala	wt/wt
	KAV_ M09	Baia di Kavala	wt/wt
	KAV_ M10	Baia di Kavala	wt/wt
	KAV_ M12	Baia di Kavala	wt/wt
	KAV_ M14	Baia di Kavala	wt/wt
	KAV_ M17	Baia di Kavala	wt/wt
	KAV_ M21	Baia di Kavala	wt/wt
TIR femmine	ORB_F14	Laguna di Orbetello (GR)	wt/m1
	ORB_F22	Laguna di Orbetello (GR)	wt/wt
	ORB_F23	Laguna di Orbetello (GR)	wt/wt
	ORB_F24	Laguna di Orbetello (GR)	wt/wt
	ORB_F25	Laguna di Orbetello (GR)	wt/wt
	CAB_F16	Laguna di Cabras (OR)	wt/wt
	CAB_F22	Laguna di Cabras (OR)	wt/m1
	CAB_F46	Laguna di Cabras (OR)	wt/wt
	CAB_F66	Laguna di Cabras (OR)	wt/wt
	CAB_F73	Laguna di Cabras (OR)	wt/wt
	TOR_F111	Laguna di Tortolì (NU)	wt/wt
	TOR_F114	Laguna di Tortolì (NU)	wt/m1
	TOR_F116	Laguna di Tortolì (NU)	wt/wt
	TOR_F126	Laguna di Tortolì (NU)	wt/m1
	TOR_F128	Laguna di Tortolì (NU)	wt/wt
	TOR_F143	Laguna di Tortolì (NU)	wt/wt
TIR maschi conformi	ORB_M04	Laguna di Orbetello (GR)	wt/m1
	ORB_M06	Laguna di Orbetello (GR)	wt/m1
	CAB_M03	Laguna di Cabras (OR)	wt/m1
	CAB_M05	Laguna di Cabras (OR)	wt/m1
	CAB_M06	Laguna di Cabras (OR)	wt/m1
	CAB_M33	Laguna di Cabras (OR)	wt/m1
	CAB_M34	Laguna di Cabras (OR)	wt/m1
	TOR_M172	Laguna di Tortolì (NU)	wt/m1
	TOR_M173	Laguna di Tortolì (NU)	wt/m1
	TOR_M174	Laguna di Tortolì (NU)	wt/m1
	ORB_M19	Laguna di Orbetello (GR)	wt/wt

TIR maschi non conformi	ORB_M21	Laguna di Orbetello (GR)	wt/wt
	CAB_M61	Laguna di Cabras (OR)	wt/wt
	CAB_M62	Laguna di Cabras (OR)	wt/wt
	CAB_M63	Laguna di Cabras (OR)	wt/wt
	CAB_M80	Laguna di Cabras (OR)	wt/wt
	CAB_M83	Laguna di Cabras (OR)	wt/wt
	TOR_M175	Laguna di Tortolì (NU)	wt/wt
	TOR_M195	Laguna di Tortolì (NU)	wt/wt
	TOR_M196	Laguna di Tortolì (NU)	wt/wt

3.2. Estrazione del DNA tramite PureLink® Genomic DNA Kit

Inizialmente è stato estratto il DNA dai campioni da analizzare utilizzando un apposito kit, chiamato PureLink® Genomic DNA Kit.

Il kit permette di eseguire una rapida ed efficiente purificazione del DNA genomico. Il DNA isolato è di 20-50 kb ed è adatto per la preparazione di librerie genomiche NGS, per PCR, digestione con enzimi di restrizione e *Southern Blotting*.

Il kit è basato sul legame selettivo del DNA con la membrana di silice in presenza di sali caotropici. Il tessuto viene digerito con un buffer e proteinasi K a 55°C; i residui di RNA vengono rimossi mediante una digestione con RNasi A. Il DNA si lega alla membrana a base di silice presente nella colonna e le impurità vengono rimosse mediante due lavaggi con Wash Buffers. Infine il DNA genomico viene eluito con un tampone di eluizione a basso contenuto di sali.

3.2.1. Preparing lysates

Per la lisi sono stati prelevati circa 25 mg di tessuto da ciascun campione e sono stati posti in tubini sterili da 1,5 ml; a ciascun campione sono stati aggiunti 180 µl di PureLink® Genomic Digestion Buffer e 20 µl di proteinasi K. Il composto ottenuto è stato incubato per circa 1 ora a 55°C, fino al raggiungimento della completa lisi dei tessuti. Al termine della lisi, i campioni sono stati centrifugati per 3 minuti alla massima velocità; il surnatante, contenente il DNA, è stato trasferito in nuovi tubini mentre il pellet è stato eliminato. Successivamente sono stati aggiunti 20 µl di RNasi A per ciascun campione e ogni soluzione è stata mescolata attraverso l'utilizzo del Vortex. Successivamente i campioni sono stati messi ad incubare per 10 minuti così da far agire l'RNasi A, necessaria per degradare eventuali residui di RNA che andrebbero a contaminare il DNA. Infine sono stati aggiunti 200 µl di PureLink® Genomic Lysis/Binding Buffer e 200 µl di etanolo.

3.2.2. Binding DNA

Il materiale lisato così ottenuto è stato trasferito all'interno di PureLink® Spin Column con Collection Tube e centrifugato per 1 minuto a 10.000 rpm. In questo modo il DNA si lega alla membrana presente nella colonna e il contenuto dei tubini, non necessario in quanto costituito da tutto ciò che non è trattenuto dalla membrana, viene scartato. Le Spin Column contenenti il DNA sono poi state trasferite in nuovi Collection Tube.

3.2.3. Washing DNA

Successivamente sono stati eseguiti due lavaggi utilizzando, per ogni campione, 500 µl di due diversi Wash Buffer.

- Il Wash Buffer 1 è stato preparando utilizzando 200 µl di Genomic Wash Buffer 1 concentrato e 300 µl di etanolo per ciascun campione; la soluzione, dopo essere stata mescolata, è stata aggiunta ai campioni che sono poi stati centrifugati per 1 minuto a 10.000 rpm. Il contenuto dei tubini è stato scartato e le Spin Column contenenti il DNA sono state trasferite in nuovi Collection Tube.
- Il Wash Buffer 2 è stato preparando utilizzando 150 µl di Genomic Wash Buffer 2 concentrato e 350 µl di etanolo per ciascun campione; la soluzione, dopo essere stata mescolata, è stata aggiunta ai campioni che sono poi stati centrifugati per 3 minuti alla massima velocità. Il contenuto dei tubini è stato scartato e le Spin Column contenenti il DNA sono state trasferite all'interno di tubini sterili da 1,5 ml.

3.2.4. Eluting DNA

L'eluizione del DNA è stata eseguita utilizzando 100 µl di PureLink® Genomic Elution Buffer per ciascun campione. Dopo aver atteso 1 minuto di incubazione ed aver centrifugato per 1 minuto alla massima velocità, le colonnine sono state scartate, mentre i tubini contenenti il DNA purificato sono stati mantenuti. Si è passati poi alla valutazione quantitativa e qualitativa del DNA e in seguito i campioni sono stati conservati alla temperatura di -20°C.

3.3. Valutazione qualitativa e quantitativa del DNA estratto

È stata eseguita una valutazione qualitativa e quantitativa del DNA ottenuto dalla purificazione, attraverso tre diverse metodiche. Come primo approccio è stata effettuata una valutazione dei campioni al NanoDrop, in seguito si è passati alla quantificazione attraverso il kit Qubit™ dsDNA Broad Range e infine è stata eseguita una valutazione qualitativa del DNA tramite elettroforesi su gel di agarosio.

La valutazione qualitativa e quantitativa del DNA estratto è necessaria per conoscere la quantità di DNA presente e se il DNA estratto è di buona qualità, in quanto successivamente dovrà essere utilizzato per la preparazione delle librerie genomiche per il sequenziamento.

3.3.1. Valutazione degli estratti tramite NanoDrop

Il NanoDrop è uno spettrofotometro che permette di avere informazioni sulla qualità e sulla quantità del DNA estratto. Lo strumento sfrutta la tecnologia basata sulla tensione superficiale dei liquidi. La goccia di campione che viene posizionata sull'apposita piastra di lettura, crea una colonna di liquido che sarà attraversata da un raggio di luce emesso da una fonte luminosa. L'intensità della luce prima e dopo il passaggio attraverso la soluzione viene misurata da un foto-detettore e viene utilizzata per calcolare l'assorbanza della sostanza. Per calcolare la concentrazione del DNA, viene sfruttata la legge di Lambert-Beer, secondo la quale l'assorbanza è direttamente proporzionale alla concentrazione della soluzione. Visto che il DNA presenta un picco di assorbanza a 260 nm, a questa lunghezza d'onda viene misurato il valore di assorbanza per calcolare la concentrazione del DNA. Vengono misurati anche i valori di assorbanza a 280 nm e a 230 nm per valutare l'eventuale presenza di proteine, carboidrati, fenoli o altri contaminanti.

Per la valutazione dei campioni, lo strumento è stato impostato per la lettura del DNA. Prima di eseguire la quantificazione è necessario pulire lo strumento con acqua sterile; successivamente sono stati aggiunti 1,5 µl di Elution Buffer, utilizzato come bianco. Infine per la misurazione sono stati inseriti 1,5 µl di ciascun campione all'interno dello strumento.

Avvenuta la lettura, il NanoDrop restituisce diversi valori: l'assorbanza a 280 nm e a 230 nm, i rapporti 260/280 e 260/230, che consentono di definire la presenza di contaminanti e la quantità di DNA, espressa in ng/µl. Il rapporto 260/280 riguarda la contaminazione da proteine, fenoli o altri contaminanti; se il DNA è di buona qualità il valore dovrebbe risultare vicino a 2. Il rapporto 260/230, invece, dà un'indicazione sulla presenza di contaminanti come carboidrati o fenoli; un valore compreso tra 1,8 e 2,2 è indice di DNA di buona qualità.

3.3.2. Valutazione degli estratti tramite Qubit™ dsDNA Broad Range

Il kit di quantificazione per DNA Qubit™ dsDNA Broad Range permette di fare una valutazione quantitativa del DNA presente nei campioni. Questo metodo si basa sull'utilizzo di un fluorimetro (QuBit™) capace di quantificare accuratamente

concentrazioni di DNA comprese tra 0,2 e 100 ng/µL. Lo strumento rileva la presenza del *dye*, un colorante che si lega al DNA e, quando viene eccitato da un raggio laser, emette fluorescenza. La misura della fluorescenza è proporzionale alla quantità di DNA; quindi per ogni campione, viene rilevata una diversa fluorescenza e, grazie alla lettura di due standard contenenti concentrazioni note di DNA, lo strumento calcola la concentrazione dei campioni (in ng/µl), a partire dal volume di campione inserito.

Tutte le operazioni necessarie per la lettura dei campioni al Qubit™ devono essere svolte in un ambiente poco luminoso in quanto il fluoroforo Qubit™ è fotosensibile, quindi la luce potrebbe compromettere la sua funzionalità.

La soluzione di lavoro Qubit™ è stata preparata con 199 µl di buffer Qubit™ e 1 µl di fluoroforo Qubit™ (rapporto 1:200) per ogni campione da esaminare e per due standard, contenenti concentrazioni note di DNA e necessari per la calibrazione dello strumento.

In seguito sono stati preparati gli standard inserendo 190 µl di soluzione di lavoro (contenente buffer e fluoroforo) e 10 µl di standard 1 (corrispondente a 0 ng/µl) in un Qubit™ Assay Tube e 190 µl di soluzione di lavoro e 10 µl di standard 2 (corrispondente a 10 ng/µl) in un altro Qubit™ Assay Tube. Per la quantificazione, invece, sono stati utilizzati 198 µl di soluzione di lavoro e 2 µl di ciascun campione di DNA da analizzare. Le miscele così composte sono state mescolate tramite Vortex ed incubate a temperatura ambiente per almeno 2 minuti per permettere alla soluzione contenente il fluoroforo di legarsi al DNA.

Si è passati infine alla lettura prima degli standard e successivamente dei campioni, dopo aver impostato lo strumento per la lettura di 2 µl di DNA.

3.3.3. Valutazione degli estratti tramite elettroforesi su gel di agarosio

L'elettroforesi su gel di agarosio è una metodica per l'analisi del DNA che consente di valutarne in modo approssimativo la concentrazione, l'integrità ed il peso molecolare. Consiste nel caricare i campioni di DNA all'interno di un gel e sottoporlo ad un campo elettrico; il DNA, essendo carico negativamente a causa dei gruppi fosfato che fanno parte della sua struttura, se sottoposto ad un campo elettrico migra verso il polo positivo. La velocità della migrazione dipende dal proprio peso molecolare: frammenti corti migrano più velocemente di frammenti lunghi.

Con l'elettroforesi è possibile valutare l'integrità di un campione di DNA; se questo è di buona qualità, al termine della corsa elettroforetica, apparirà una banda nitida, cioè il campione risulterà costituito da frammenti di lunghezza omogenea. Diversamente, se il

DNA ottenuto dall'estrazione non è di buona qualità, ovvero contiene DNA degradato, al termine della corsa si avrà una banda non nitida che apparirà come uno *smear* perché i frammenti di DNA di diversa lunghezza saranno migrati diversamente: in basso i frammenti a basso peso molecolare e in alto quelli ad alto peso molecolare.

Per la valutazione del DNA estratto è stato utilizzato un gel di agarosio allo 0,8% (per la separazione di frammenti ad alto peso molecolare). Il gel è stato preparato utilizzando 50 ml di buffer TAE 1X e 0,4 g di agarosio; il composto è poi stato riscaldato con l'obiettivo di sciogliere l'agarosio ed ottenere una soluzione omogenea. Successivamente sono stati aggiunti 5 µl di SybrSafe, un intercalante del DNA, che sono stati mescolati al composto. In seguito il gel è stato colato all'interno del vassoio precedentemente preparato con l'apposito pettine, necessario per la formazione dei pozzetti. Una volta che il gel si è solidificato, è stato tolto il pettine e il gel è stato posto all'interno dell'apposita vaschetta contenente una soluzione tampone che permette la conduzione della corrente elettrica.

Si è passati poi al caricamento dei campioni di DNA all'interno dei pozzetti. Per ciascun campione di DNA, è stata preparata la soluzione da caricare nel gel con 5 µl di DNA e 5 µl di *loading dye* 2X, utile per visualizzare la migrazione del DNA durante la corsa e per agevolare la precipitazione del campione sul fondo del pozzetto, grazie alla presenza di glicerolo. Nel primo pozzetto è stato caricato il marcatore di peso molecolare λHindIII, grazie al quale è possibile valutare le dimensioni delle bande.

Dopo il caricamento dei campioni, la cella contenente il gel è stata collegata al generatore di corrente, impostando un voltaggio di 90 volt. Attesi circa 30 minuti la corsa è stata interrotta e si è passati alla visualizzazione del gel tramite l'esposizione a raggi UV.

3.4. Preparazione delle librerie genomiche tramite il protocollo Illumina DNA Prep Tagmentation

Per la preparazione delle librerie genomiche è stato utilizzato il kit Illumina DNA Prep Tagmentation con volumi di reagenti dimezzati rispetto a quanto indicato dal protocollo.

3.4.1. Tagment Genomic DNA

La prima fase del protocollo prevede l'utilizzo di trasposoni legati alle microsfere (Bead-Linked Transposomes, BLT) per la tagmentazione del DNA, una reazione che permette di frammentare il DNA e di marcarlo con gli adattatori per il sequenziamento.

Per svolgere questa operazione è stato prelevato, per ciascun campione, un volume di DNA tale da totalizzare 150 ng di templato di partenza ed è stata aggiunta acqua *nuclease-free* per raggiungere il volume totale di 15 µl. È stata poi preparata una soluzione contenente 5,5 µl di BLT e 5,5 µl di Tagmentation Buffer 1 (TB1) per ciascun campione; questa è stata mescolata tramite Vortex e successivamente sono stati aggiunti 11 µl di soluzione a ciascun campione. I campioni sono stati quindi incubati a 55°C per 15 minuti per far avvenire la reazione di tagmentazione.

3.4.2. Post Tagmentation Cleanup

Attesi i 15 minuti, sono stati aggiunti 5 µl di Tagment Stop Buffer (TSB) a ciascun campione e le soluzioni sono state ben mescolate per risospendere le biglie. L'aggiunta di TSB permette di interrompere la reazione di tagmentazione, evitando una eccessiva frammentazione del DNA. Per una buona riuscita delle librerie, questo passaggio deve essere svolto il più velocemente possibile, evitando di superare i 15 minuti di incubazione, altrimenti il DNA potrebbe risultare troppo frammentato. In seguito all'aggiunta di TSB i campioni sono stati incubati a 37°C per 15 minuti.

Il passaggio successivo prevede tre lavaggi con Tagment Wash Buffer (TWB), allo scopo di eliminare i reagenti della reazione di tagmentazione e lavare il complesso "trasposone-DNA". Per fare questo le provette contenenti i campioni sono state posizionate su un supporto magnetico che consente di attrarre le BLT a cui è legato il DNA e di eliminare ciò che non deve essere mantenuto nei passaggi successivi.

I lavaggi sono stati eseguiti utilizzando 100 µl di TWB a campione per ciascun lavaggio; le seguenti operazioni sono state ripetute per tre volte:

- rimuovere i campioni dal supporto magnetico e aggiungere 100 µl di TWB a ciascun campione
- mescolare bene fino a che le biglie non risultano completamente risospese
- posizionare i campioni sul supporto magnetico ed attendere circa 3 minuti per fare aderire le biglie al magnete
- rimuovere il surnatante ed eliminarlo, facendo attenzione a non prelevare le biglie.

Dopo aver concluso l'ultimo lavaggio, i campioni sono stati tolti dal supporto magnetico e si è passati alla fase di amplificazione del DNA tramite PCR.

3.4.3. Amplify Tagmented DNA

Questo passaggio prevede l'esecuzione della reazione di PCR, che permette di amplificare il DNA, e l'aggiunta degli Index 1 (i7) e Index 2 (i5), che serviranno per il pooling di più campioni in una singola corsa di sequenziamento.

La miscela di reazione per la PCR è stata preparata utilizzando 11 µl di Enhanced PCR Mix (EPM) e 11 µl di acqua *nuclease-free* per ciascun campione; la soluzione è stata vortexata e aggiunta ai campioni (22 µl per ciascun campione) che sono poi stati mescolati per risospendere le biglie. Successivamente sono stati aggiunti 5 µl di index i7 e i5 pre-miscelati, i campioni sono stati mescolati con gli index ed è stata avviata la PCR, precedentemente impostata secondo il seguente programma:

- 68°C per 3 minuti
- 98°C per 3 minuti
- 5 cicli di:
    - 98°C per 45 secondi
    - 62°C per 30 secondi
    - 68°C per 2 minuti
- 68°C per 1 minuto

Una volta conclusa la reazione di PCR, si è passati alla fase di purificazione delle librerie.

3.4.4. Clean Up Libraries

La fase di purificazione delle librerie, eseguita utilizzando le Sample Purification Beads (SPB), permette di selezionare soltanto i frammenti di DNA della lunghezza desiderata. Attraverso due step successivi vengono quindi rimosse prima le molecole di DNA di dimensioni superiori ad un determinato valore e poi i frammenti troppo corti.

Per svolgere questa fase è stato nuovamente utilizzato un supporto magnetico, sopra al quale sono stati posizionati i prodotti di PCR. Attesi circa 5 minuti, necessari per far avvenire l'attrazione tra le biglie e il supporto magnetico, sono stati trasferiti 23 µl di surnatante in una nuova piastra e a questi sono stati aggiunti 62 µl di acqua *nuclease-free* per raggiungere un volume totale di 85 µl per ogni campione. Sono stati poi addizionati, per ciascun pozzetto, 45 µl di SPB. Dopo aver mescolato e incubato per 5 minuti a temperatura ambiente su un agitatore per ottenere una soluzione omogenea, la piastra è stata trasferita sul supporto magnetico e sono stati attesi circa 5 minuti. In questo modo è stato possibile prelevare il surnatante, contenente i frammenti di DNA di dimensioni minori, ed eliminare i frammenti di dimensioni maggiori che si sono legati alle biglie. Sono quindi stati trasferiti 125 µl di surnatante per ciascun campione in un nuovo

pozzetto e sono stati aggiunti 15 µl di SPB. Dopo aver mescolato e incubato per 5 minuti a temperatura ambiente su un agitatore fino ad ottenere una soluzione omogenea, la piastra è stata trasferita sul supporto magnetico e il surnatante, contenente i frammenti di DNA di piccole dimensioni, è stato eliminato. In questo modo è stato mantenuto soltanto il DNA delle dimensioni di interesse, legato alle biglie.

Successivamente sono stati effettuati due lavaggi con etanolo 80%; a questo scopo, sono stati ripetuti i seguenti passaggi per due volte:

- mantenendo la piastra sul supporto magnetico, aggiungere 200 µl di etanolo 80% a ciascun campione
- incubare per 30 secondi
- rimuovere il surnatante, facendo attenzione a non prelevare le biglie

Una volta conclusi i due lavaggi, sono stati attesi circa 5 minuti per far evaporare l'etanolo residuo. Infine la piastra è stata rimossa dal supporto magnetico e sono stati aggiunti 25 µl di Resuspension Buffer (RSB) per ciascun campione, così da eluire il DNA dalle biglie.

3.5. Valutazione quantitativa e qualitativa delle librerie genomiche

In seguito alla creazione delle librerie genomiche, si è passati alla valutazione quantitativa del DNA, necessaria per accertarsi della buona riuscita delle librerie e per definire la loro concentrazione. Per fare ciò è stato utilizzato il kit di quantificazione specifico per DNA Qubit™ dsDNA High Sensitivity.

Per ciascuna libreria, è stata poi misurata la qualità dei campioni tramite lo strumento Bioanalyzer.

3.5.1. Valutazione tramite Qubit™ dsDNA High Sensitivity

Il kit di quantificazione per DNA Qubit™ dsDNA High Sensitivity prevede l'utilizzo dello stesso strumento e dello stesso procedimento usato per la quantificazione del DNA fatta in seguito all'estrazione, con la differenza che il kit High Sensitivity è capace di quantificare accuratamente concentrazioni di DNA comprese tra 0,01 e 10 ng/µL.

3.5.2. Valutazione tramite Bioanalyzer High Sensitivity DNA Assay

La metodica Bioanalyzer 2100 (Agilent technologies 2100) prevede la corsa dei campioni in un chip miniaturizzato contente del gel. Il chip ha 16 pozzetti, di cui 11 destinati ai campioni, collegati tra loro da microcanali di vetro all'interno dei quali andrà inserito il *Gel-Dye* che permette la separazione del campione sulla base del peso molecolare. Il principio è lo stesso di una corsa elettroforetica classica ma con tempi ridotti e maggiore

sensibilità. La rilevazione dei frammenti di DNA avviene attraverso una fluorescenza laser-indotta che sfrutta un colorante fluorescente intercalante del DNA.

Per la valutazione dei campioni tramite Bioanalyzer sono state eseguite le seguenti operazioni:

- Sono stati caricati 9 µL di *Gel-Dye Mix* nel fondo del pozzetto indicato con la lettera G cerchiata di nero, cercando di non formare bolle d'aria, che potrebbero alterare il risultato.
  Il *Gel-Dye Mix*, mantenuto a 4°C, deve essere posto a temperatura ambiente per almeno 30 minuti prima di essere utilizzato; va inoltre mantenuto al riparo dalla luce a causa della sua fotosensibilità.
- Dopo aver verificato che lo stantuffo della siringa sia posizionato esattamente su 1 ml, è stata chiusa la chip printing station. Dopo 1 minuto lo stantuffo è stato rilasciato. Una volta rilasciato, lo stantuffo risale autonomamente fino a raggiungere i 0,7 ml circa e poi, dopo aver atteso qualche secondo, è stato sollevato manualmente per portarlo alla posizione iniziale di 1 ml. Queste operazioni sono necessarie per distribuire il gel in maniera omogenea all'interno dei microcanali.
- Successivamente sono stati caricati 9 µL di *Gel-Dye Mix* nei rimanenti pozzetti contrassegnati con la lettera G, mentre in tutti gli altri pozzetti sono stati posti 5 µL di *Marker*. È necessario assicurarsi di non lasciare alcun pozzetto vuoto, altrimenti lo strumento non funzionerà.
- Nel pozzetto contenente il simbolo di una scala è stato poi aggiunto 1 µL di *Ladder*, un marcatore di peso molecolare che permette di definire la lunghezza dei frammenti di DNA. In seguito si è passati al caricamento dei campioni: negli 11 pozzetti destinati ai campioni è stato aggiunto 1 µL di DNA.
- Concluso il caricamento, il chip è stato agitato tramite vortex per 1 minuto a 2.400 rpm. Dopo aver verificato che nel chip non fossero presenti bolle d'aria, si è passati al suo inserimento all'interno dello strumento per la corsa.

Figura 7. *Chip utilizzato nella metodica Bioanalyzer High Sensitivity DNA Assay*

3.6. Sequenziamento Illumina

Le librerie preparate sono state inviate presso il Norwegian Sequencing Centre per effettuare il sequenziamento attraverso la tecnologia Illumina.

Il sequenziamento Illumina prevede tre passaggi principali:

- Preparazione delle librerie
  Il DNA viene frammentato in segmenti di una data lunghezza a cui verranno aggiunti adattatori specifici ad entrambe le estremità. In seguito viene eseguita un'amplificazione delle sequenze, in modo da fornire un segnale sufficiente per il sequenziamento.
- Generazione di cluster
  I frammenti vengono clonati in migliaia di copie identiche. Per generare i cluster, la libreria viene caricata in una *flowcell*, cioè un vetrino con diverse corsie la cui superficie è ricoperta da oligonucleotidi complementari agli adattatori presenti sulle librerie. Una volta legata alla *flowcell*, la libreria viene estesa dalla polimerasi per ottenere cluster con migliaia di copie della molecola iniziale.
- Sequenziamento
  Inizia con l'introduzione del primo primer di sequenziamento nella *flowcell* e la sua ibridazione all'adattatore. Ad ogni ciclo poi, vengono aggiunti i nucleotidi marcati mediante fluorofori che competono per essere addizionati alla catena in crescita. Soltanto il nucleotide complementare alla sequenza modello viene incorporato, mentre gli altri vengono lavati via. Dopo l'aggiunta di ogni nucleotide, i cluster vengono eccitati da una sorgente luminosa e ciascuna base emette ad una specifica lunghezza d'onda (https://www.illumina.com/). Esistono tre diverse

metodologie per differenziare le basi; a seconda del sistema utilizzato, si può avere un sequenziamento con chimica a

- quattro canali: vengono utilizzati quattro coloranti fluorescenti, uno per ogni base, e quattro immagini per ciclo di sequenziamento
- due canali: vengono utilizzati due coloranti fluorescenti e due immagini per ciclo di sequenziamento per codificare i dati per le quattro basi del DNA. Sono presenti un'immagine dal canale rosso e un'immagine dal canale verde; una mancata identificazione viene indicata con una N
- un canale: utilizza la tecnologia CMOS per l'identificazione delle basi utilizzando due immagini per ciclo di sequenziamento. I nucleotidi vengono identificati tramite l'analisi dei diversi schemi di emissione per ogni base attraverso le due immagini. L'adenina presenta una marcatura rimovibile ed è marcata solo nella prima immagine. La citosina presenta un gruppo che si lega a una marcatura ed è marcata solo nella seconda immagine. La timina presenta una marcatura fluorescente ed è dunque marcata in entrambe le immagini, mentre la guanina è sempre scura (non marcata).

Nel presente lavoro di tesi, lo strumento utilizzato per il sequenziamento Novaseq6000 utilizza la chimica a due canali.

È stata utilizzata una lane di una *flowcell* NovaSeq S4. Ogni *flowcell* NovaSeq S4 contiene quattro lane e fornisce fino a 3 TB di dati. Una lane di *flowcell* NovaSeq S4, sequenziata in modalità 150 PE, produce in media 750 Gb di dati. Nel caso di 32 librerie genomiche di *Mugil cephalus*, il cui genoma ha una dimensione stimata di 0,8 Gb (Raymond, et al. 2022), questo permette di ottenere un *coverage* per libreria di almeno 25X.

3.7. Analisi dei dati di sequenziamento

Una volta eseguito il sequenziamento, i dati vengono restituiti in formato FASTQ, un file di testo contenente la sequenza nucleotidica e il punteggio di qualità di ogni base.

Si procede quindi con le analisi dei dati di sequenziamento tramite diversi software.

3.7.1. Valutazione della qualità delle sequenze attraverso il software FastQC

Il primo passaggio di lavoro che viene eseguito è la valutazione della qualità delle sequenze grezze attraverso il software FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) che restituisce, per ciascun campione, un file HTML contenente grafici e informazioni relative alla qualità delle sequenze.

Il comando che è stato utilizzato per eseguire questa operazione è il seguente:

```
#quality control checks
fastqc *fastq.gz
```

I principali grafici che vengono creati dal software sono:

- *Per base sequence quality*
  Il grafico permette di visualizzare la qualità, in termini di *Phred score*, delle *reads* dei campioni. Per ogni posizione viene disegnato un diagramma di tipo *BoxWhisker*, che consiste in una rappresentazione grafica utilizzata per descrivere la distribuzione di un campione.
- *Per sequence quality scores*
  Viene calcolata dal software la qualità media di ciascuna *read* e, attraverso il grafico, si può visualizzarne la distribuzione.
- *Per base N content*
  Il grafico permette di valutare la quantità delle basi non identificate, indicate con la lettera N. Per poter definire i dati di buona qualità, il contenuto di N dovrebbe essere <5%.
- *Sequence length distribution*
  Il grafico riporta la distribuzione della lunghezza delle sequenze generate in seguito al sequenziamento.
- *Adapter content*
  Il grafico restituisce una stima della presenza di adattatori all'interno della sequenza.

3.7.2. *Trimming* delle sequenze grezze attraverso il software Trim Galore!

Successivamente è stato eseguito il *trimming* delle sequenze grezze attraverso il software Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Il *trimming* è un processo che consente di eliminare gli adattatori e le porzioni di bassa qualità presenti nelle sequenze.

Per eseguire questa operazione è stato utilizzato un ciclo for, cioè un comando che permette di eseguire la medesima analisi per una lista di campioni. Nel nostro caso è stato creato un file di testo chiamato 'Sample_list.txt' contenente la lista dei campioni da processare.

```
for s in $(cat Sample_list.txt);
do
trim_galore -q 25 --fastqc  --length 70 --paired --clip_R1 5 --clip_R2 5 --nextera --cores 6 -o ~/MuCe/trimmed \\
--basename ${s} ~/MuCe/raw_reads/${s}_L002_R1_001.fastq.gz ~/MuCe/raw_reads/${s}_L002_R2_001.fastq.gz
done
```

Sono stati impostati vari parametri per ottenere delle sequenze di buona qualità.

- Il parametro -q permette di impostare la soglia minima di qualità di ciascuna base; nel nostro caso è stato impostato un *Phred score* di 25, che corrisponde ad una probabilità pari a 0,00316 che quella base venga letta in modo errato.
- Il parametro --length serve per impostare la minima lunghezza delle *reads* da mantenere in seguito al trimming.
- L'opzione --paired permette di eseguire il *trimming* per i file accoppiati. Per superare il test di convalida, entrambe le sequenze di una coppia di sequenze devono essere della lunghezza minima (determinata dal parametro --length).
- Attraverso il comando --clip_R1 e --clip_R2 sono state rimosse dall'estremità 5' le prime 5 basi sia per il *forward* (R1) che per il *reverse* (R2) per rimuovere il rumore di fondo presente all'inizio di una lettura.
- L'opzione --nextera consente di riconoscere la sequenza corretta degli adattatori per rimuoverli.
- Il parametro --cores serve per definire il numero di *core* (processori) utilizzati.
- Il comando -o viene utilizzato per indicare il percorso della cartella su cui salvare i file di output.

3.7.3. *Mapping* delle sequenze sul genoma di riferimento attraverso il software BWA

In seguito i file di sequenza sono stati mappati sul genoma di riferimento di *M. cephalus* tramite il software BWA (Li e Durbin 2010).

BWA è un software per il *mapping* di sequenze composto da tre algoritmi: BWA-backtrack, BWA-SW e BWA-MEM. Il primo algoritmo è progettato per letture di sequenze Illumina fino a 100 bp, mentre gli altri due per sequenze più lunghe (da 70 bp a 1 Mbp). BWA-MEM è generalmente consigliato per *query* di alta qualità poiché risulta essere più veloce e più preciso in confronto alle altre due.

Il primo passaggio necessario per lavorare con BWA consiste nell'indicizzare il genoma di riferimento, per poi riuscire ad allineare i dati. Per fare questo è stato utilizzato il seguente comando:

```
#GENOME INDEXING for BWA
bwa index ./Genome/Mcephalus_assembly.fa
```

È stato poi utilizzato l'algoritmo BWA-MEM, che funziona allineando i segmenti con le massime corrispondenze esatte (M*aximal Exact Matches*, MEMs) confrontandole con il genoma di riferimento. L'algoritmo BWA-MEM produce allineamenti locali. Può produrre allineamenti multipli per le diverse parti di una sequenza; questa caratteristica è importante per le sequenze lunghe. Alcuni programmi però, ad esempio *Picard's markDuplicates*, non funzionano con allineamenti divisi, quindi bisogna usare l'opzione '-M' per contrassegnare gli split hit più brevi come secondari.

Il comando utilizzato è:

```
for s in $(cat Sample_list.txt);
do
bwa mem ./genome/Mcephalus_assembly.fa ./trimmed/${s}_val_1.fq.gz ./trimmed/${s}_val_2.fq.gz \\
-t 15 -c 1 -M | samtools sort -@15 -o ./BWA/${s}.bam -
done
```

- Il parametro -t serve per impostare il numero di *core* utilizzati dal software.
- L'opzione -c permette di determinare il numero massimo di match che ciascuna sequenza può avere lungo il genoma. Superata la soglia impostata, la sequenza viene scartata. Nel nostro caso, dovendo utilizzare le sequenze per successivo *SNP calling*, è stato scelta massima stringenza tenendo solo le *read* cha avessero un unico match in tutto il genoma.
- -M serve per contrassegnare gli *split hit* più brevi come secondari (necessario per la compatibilità con Picard).

Durante la creazione dell'output di BWA-MEM viene utilizzato il software SAMtools (Danecek, Bonfield, et al. 2021) per convertire i file dal formato SAM al formato BAM. SAM è un formato standard per la memorizzazione di allineamenti di sequenze di grandi dimensioni, mentre il formato BAM è la forma binaria di SAM. SAMtools consiste un insieme di programmi che permettono di interagire con dati di sequenziamento di alta qualità e serve per manipolare gli allineamenti in formato SAM/BAM (ordinamento, unione, indicizzazione).

Il comando sort viene utilizzato per ordinare il contenuto del file per coordinate di allineamento e attraverso - @ viene impostato il numero di core da utilizzare.

3.7.4. Creazione dei *read group* e rimozione delle letture duplicate attraverso lo strumento Picard

Per la creazione dei *read group* e la marcatura e rimozione delle letture duplicate è stato utilizzato Picard (https://broadinstitute.github.io/picard/), uno strumento composto da una serie di linee di comando utili a elaborare i dati delle sequenze *high-throughput* (HTS).

Per assegnare ad ogni campione un *read group* (un insieme di letture generate da una singola corsa di uno strumento di sequenziamento) è stato utilizzato il seguente comando:

```
#READ GROUP & REMOVE DUPLICATES & SORTING
# $PICARD -> pathway for  picard.jar
for s in $(cat Sample_list.txt);
do
java -jar $PICARD AddOrReplaceReadGroups INPUT= ./BWA/${s}.bam OUTPUT= ./BWA/${s}_RG.bam \\
VALIDATION_STRINGENCY=SILENT RGID=${s} RGLB=DNA RGPL=Illumina RGPU=01 RGSM=${s}
done
```

- AddOrReplaceReadGroups è un comando che consente di sostituire tutti i *read groups* presenti nel file di input con un nuovo singolo *read group* e di assegnare tutte le *reads* a questo *read group* nel file di output.
- L'opzione INPUT serve per definire quali sono i file di input (che finiscono con '.bam') e per indicare il percorso della cartella all'interno della quale questi si trovano.
- L'opzione OUTPUT serve per definire quali saranno i file di output (che finiranno con 'RG.bam') e per indicare il percorso della cartella all'interno della quale verranno messi i file.
- L'impostazione VALIDATION_STRINGENCY=SILENT può migliorare le prestazioni durante l'elaborazione di un file BAM in cui non è necessario decodificare i dati di lunghezza variabile.
- RGID (Read Group ID): per assegnare un ID ad ogni *read group*.
- RGLB (Read Group library): per indicare che stiamo lavorando con sequenze di DNA.
- RGPL (*Read Group Platform*): per indicare che la libreria è stata creata con la piattaforma "Illumina".
- RGPU (*Read Group Platform Unit*): definizione specifica per un gruppo di letture.
- RGSM (*Read Group Sample Name*): per impostare il nome dei campioni a cui verrà assegnato il *read group*.

In seguito, è stato utilizzato lo strumento *MarkDuplicates* di Picard, che permette di individuare e contrassegnare le letture duplicate in un file BAM o SAM; le letture duplicate sono originate da un singolo frammento di DNA e possono verificarsi durante la preparazione del campione, ad esempio la costruzione di librerie mediante PCR. Possono anche derivare da duplicati ottici, ossia quando un singolo cluster di amplificazione viene erroneamente rilevato come più cluster dal sensore ottico dello strumento di sequenziamento.

Lo strumento *MarkDuplicates* funziona confrontando le sequenze nelle posizioni 5' sia delle *reads* che delle *reads-paired* in un file SAM o BAM. L'output è un nuovo file SAM o BAM, in cui i duplicati sono stati identificati e contrassegnati con il valore esadecimale di 0x0400, che corrisponde a un valore decimale di 1024.

Il comando che è stato eseguito è il seguente:

```
#REMOVE DUPLICATES
for s in $(cat Sample_list.txt);
do
java -jar $PICARD MarkDuplicates INPUT= ./BWA/${s}_RG.bam OUTPUT= ./BWA/${s}_RG_MD.bam  \\
M=./BWA/${s}_metrics.txt REMOVE_DUPLICATES=true CREATE_INDEX=true VALIDATION_STRINGENCY=SILENT
done
```

- L'opzione M serve per creare un file di output con le metriche di duplicazione per ogni campione.
- Il parametro REMOVE_DUPLICATES=true, se impostato su 'true', permette di non scrivere i duplicati nel file di output invece di scriverli con gli appropriati flag impostati.
- Il parametro CREATE_INDEX=true permette di creare un indice quando viene scritto il file BAM ordinato secondo le coordinate.
- L'impostazione VALIDATION_STRINGENCY=SILENT può migliorare le prestazioni durante l'elaborazione di un file BAM in cui non è necessario decodificare i dati di lunghezza variabile.

3.7.5. *SNP calling* attraverso lo strumento BCFtools

L'ultimo passaggio consiste nello *SNP calling*, ossia l'identificazione dei polimorfismi presenti in ciascun campione rispetto ad un genoma di riferimento. Questa operazione permette di ottenere un file VCF con all'interno le coordinate degli SNP identificati. Un file VCF (Variant Call Format) è un file tabulare in cui sono contenute le informazioni riguardanti gli SNP riscontrati nei campioni analizzati, tra cui la posizione in cui è stata riscontrata la variante, l'allele di riferimento presente sulla posizione specificata e l'allele alternativo, la qualità dell'allele alternativo, e il genotipo per ogni campione analizzato.

Per lo *SNP calling* è stato utilizzato il software BCFtools (H. Li 2011), che permette di lavorare con file in formato VCF o con la forma binaria, BCF.

```
##SNP calling
bcftools mpileup --threads 14 -b ./BAM_list.txt  -r contig_2337_pilon_pilon --fasta-ref ../Genome/Mcephalus_genome_reduced.fa \
-q 20  | bcftools call -cv -O z -o ./VCF/Mcephalus_FSHR.vcf.gz
```

Il primo comando che è stato utilizzato è bcftools mpileup, che genera un file VCF/BCF contenente le probabilità di genotipo per i file di allineamento.

- L'opzione --threads serve per definire il numero di processori utilizzati.

- Il parametro -b indica che verrà usato un file di testo contenente la lista dei file BAM.
- Il comando -r viene utilizzato per selezionare la regione del genoma su cui sarà effettuato lo *SNP calling*.
- L'opzione --fasta -ref indica la sequenza di riferimento che è in formato fasta.
- Tramite il parametro -q è stata impostata la qualità di *mapping* minima utilizzabile per un allineamento.

In seguito è stato eseguito il comando bcftools call, impostando i seguenti parametri:

- -cv: indica che il metodo di chiamata usato è SAMtools/BCFtools e nell'output verranno inserite solo le posizioni delle varianti.
- -O per determinare il formato del file di output, usando il valore "z" il file VCF sarà compresso.
- -o per indicare il nome del file di output e la cartella all'interno della quale verrà inserito.

Lo strumento bcftools filter serve invece per eseguire un filtraggio all'interno di file VCF/BCF, impostando un determinato valore soglia. Il comando utilizzato è il seguente:

```
bcftools filter --SnpGap 5 Mcephalus_FSHR.vcf.gz -O z -o Mcephalus_FSHR_SnpGap.vcf.gz
```

È stata usata l'opzione --SnpGap per filtrare solo gli SNP entro 5 paia di basi di un *indel* o un altro tipo di variante.

Successivamente è stato usato lo strumento VCFtools (Danecek, Auton, et al. 2011), che consiste in un insieme di funzioni utilizzate per lavorare su dati in formato VCF/BCF. Questo strumento serve principalmente per riepilogare dati, eseguire calcoli, filtrare dati e convertirli in formati differenti.

```
vcftools --gzvcf Mcephalus_FSHR_SnpGap.vcf.gz --min-alleles 2 --max-alleles 2 --remove-indels --recode \
--recode-INFO-all --stdout >Mcephalus_FSHR_final.vcf.gz
```

- L'opzione --gzvcf viene utilizzata per leggere file in formato VCF compressi.
- --min-alleles e --max-alleles permettono di includere nell'analisi solo i siti con un numero di alleli maggiore o uguale al valore impostato per "--min-alleles" e minore o uguale al valore impostato per "--max-alleles". Nel nostro caso vengono inclusi solo siti bi-allelici.
- L'opzione --remove-indels è necessaria per escludere i siti che contengono un *indel* (ogni variante che altera la lunghezza dell'allele di riferimento).

- L'opzione --recode-INFO-all: è stata utilizzata per definire un'informazione chiave da inserire nel nome dei file di output; nel nostro caso vengono mantenute tutte le informazioni dei file originali.
- Il parametro --stdout viene utilizzato per avere uno standard output.

Infine è stato utilizzato lo strumento bcftools stats che analizza file VCF o BCF e produce un file di testo contenente le statistiche del file VCF/BCF.

```
bcftools stats -s - Mcephalus_FSHR_final.vcf.gz >SNP_FSHR_final.stats.txt
```

Viene utilizzata l'opzione - s per indicare che verranno usati tutti i campioni.

3.7.6. Calcolo F_{ST} e creazione heatmap

È stato infine calcolato l'F_{ST} (*Fixation index*), un indice della distanza genetica tra due popolazioni. Il valore dell'F_{ST} può variare tra 0 e 1; un valore pari a 0 significa condivisione completa del materiale genetico tra le due popolazioni, mentre 1 indica che non c'è nessuna condivisione di materiale genetico tra le popolazioni analizzate.

L'indice è stato calcolato per valutare la differenza delle frequenze alleliche tra maschi e femmine. Il calcolo è stato effettuato sia per la popolazione del Tirreno (TIR), che per la popolazione dell'Egeo (KAV).

Gli F_{ST} sono stati calcolati attraverso il software RStudio (https://posit.co/products/open-source/rstudio/), un ambiente di sviluppo interattivo (IDE) disponibile per il linguaggio R. R è un linguaggio open source usato in genetica per l'analisi statistica.

Dopo aver effettuato il calcolo degli F_{ST}, per ogni gruppo di confronto, sono stati selezionati gli SNP con i valori più elevati. Utilizzando le posizioni degli SNP ed i genotipi degli individui, sempre attraverso RStudio, è stata realizzata una *heatmap*, ossia una rappresentazione grafica che, attraverso l'utilizzo dei colori, permette di visualizzare i genotipi delle diverse posizioni.

4. Risultati

4.1. Qualità del DNA estratto

Il DNA estratto dai campioni è stato valutato tramite NanoDrop per avere informazioni sulla purezza e sulla quantità del DNA, Qubit™ per una valutazione quantitativa precisa ed infine elettroforesi su gel di agarosio, una metodica che consente di valutare l'integrità del DNA estratto.

I risultati delle analisi effettuate attraverso le tre diverse metodiche hanno permesso di definire la buona qualità del DNA estratto dai campioni. La quantità è risultata sufficiente per eseguire la fase successiva, ossia la preparazione delle librerie genomiche per il sequenziamento.

Nella figura 8 viene riportato un esempio di risultati ottenuti dall'elettroforesi su gel di agarosio. La maggior parte dei campioni presentano una banda nitida, che significa che il DNA estratto risulta integro. Solo in alcuni campioni appare un lieve "*smear*", indice della presenza anche di DNA degradato.

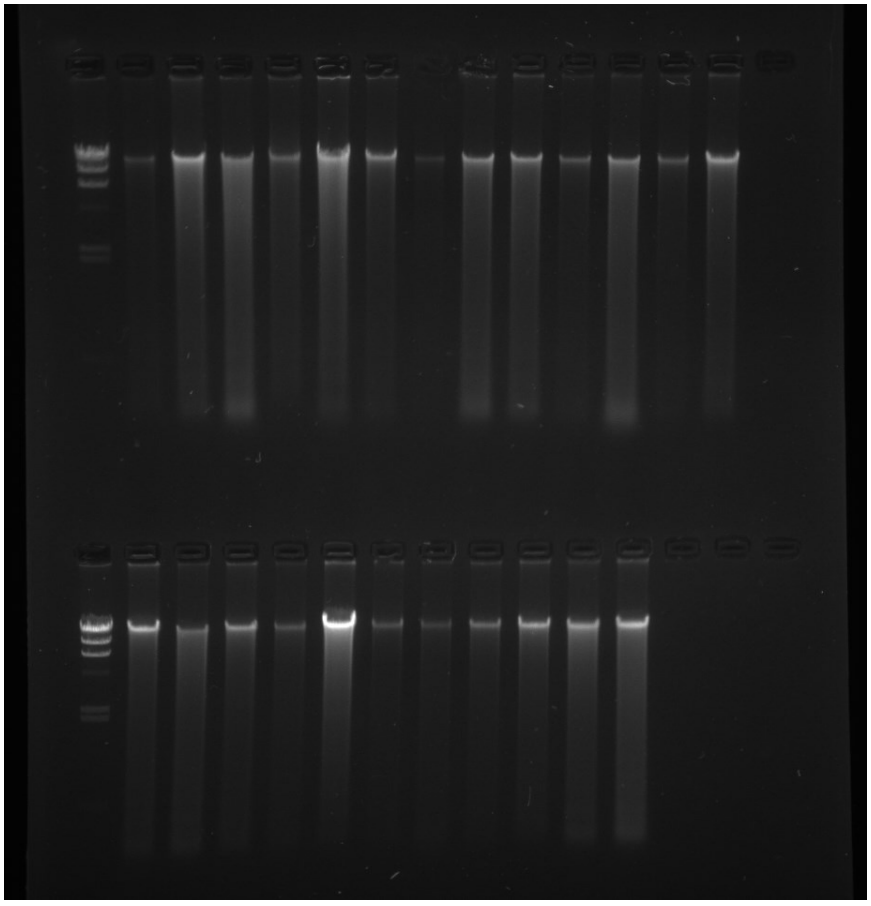


Figura 8. *Esempio di risultati dell'elettroforesi su gel di agarosio. Nel primo pozzetto di ciascuna riga è stato caricato il marcatore di peso molecolare λHindIII.*

4.2. Qualità delle librerie genomiche

La qualità delle librerie genomiche è stata valutata grazie alle analisi con Qubit™ e Bioanalyzer, due strumenti che permettono di effettuare una valutazione rispettivamente quantitativa e qualitativa del DNA. La qualità è risultata essere buona per tutte le librerie.

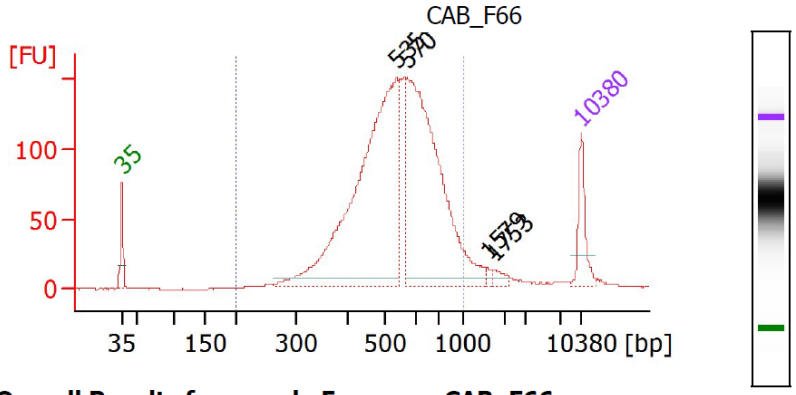
La concentrazione media delle librerie genomiche è di 5,96 ng/µl (range 2,87 – 9,33 ng/µl), corrispondente ad una molarità di 15,35 nM (range 6,84 – 22,30 nM); quest'ultimo valore risulta quindi ottimale per il sequenziamento, che richiede una molarità compresa tra i 2 e i 30 nM.

Tabella 3. *Per ciascun campione viene riportata la concentrazione del DNA estratto, la concentrazione delle librerie genomiche e la loro molarità.*

Campione	Concentrazione DNA estratto (ng/µl)	Concentrazione librerie genomiche (ng/µl)	Molarità (nM)
KAV_F02	464	4,75	9,97
KAV_F04	338	5,32	12,54
KAV_F06	63,9	6,9	20,73
KAV_F07	44,5	6,4	18,97
KAV_F09	570	6,54	19,23
KAV_F10	630	6,27	18,70
KAV_F11	224	7,42	22,30
KAV_F12	151	7,19	17,15
KAV_F21	146	6,25	20,82
KAV_F22	365	8,36	16,48
KAV_F24	111	8,82	15,99
KAV_F25	384	9,33	17,51
KAV_F27	459	7,64	17,53
KAV_F29	283	5,14	14,09
KAV_F30	342	6,01	15,52
KAV_F31	241	4,46	10,76
ORB_F14	23,3	6,32	18,02
ORB_F22	34	6,89	21,18
ORB_F23	32	6,78	21,10
ORB_F24	27,4	5,09	13,60
ORB_F25	25,4	4,02	11,07
CAB_F16	38,8	4,66	11,57
CAB_F22	50,3	6,62	18,54
CAB_F46	51,6	5,52	16,44
CAB_F66	31	2,87	6,84

CAB_F73	25,8	4,55	14,32
TOR_F111	32,4	3,97	7,87
TOR_F114	18,8	4,84	11,48
TOR_F116	66,1	5,61	11,81
TOR_F126	83,9	5,09	13,49
TOR_F128	38,5	5,24	13,30
TOR_F143	59	5,91	12,13

Il Bioanalyzer restituisce, per ogni campione, un grafico che permette di visualizzare le dimensioni dei frammenti generati e le quantità di tali frammenti. Il software dello strumento calcola inoltre una stima della dimensione media della libreria e la sua molarità. Nell'esempio riportato in figura 9, si può osservare un picco principale che indica che la maggior parte dei frammenti hanno una dimensione di circa 500 pb, con una distribuzione complessiva che va da 300 a 1000 bp. I due picchi più piccoli (a 35 e 10.380 pb) rappresentano invece i marcatori di peso molecolare, necessari per stimare le dimensioni dei frammenti. Tramite l'analisi con il Bioanalyzer si ottiene inoltre una rappresentazione grafica (figura 10) che permette di visualizzare le dimensioni dei frammenti ottenuti. La banda più scura indica la presenza di una maggiore quantità di frammenti di una dimensione compresa tra 400 e 700 pb.



Overall Results for sample 5 : CAB_F66

Number of peaks found:	4
Noise:	0.2
Corr. Area 1:	2,534.1

Region table for sample 5 : CAB_F66

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,000	2,534.1	92	548	24.3	2,271.83	6,835.1	

Figura 9. Esempio di grafico ottenuto dall'analisi tramite Bioanalyzer (campione CAB_F66).

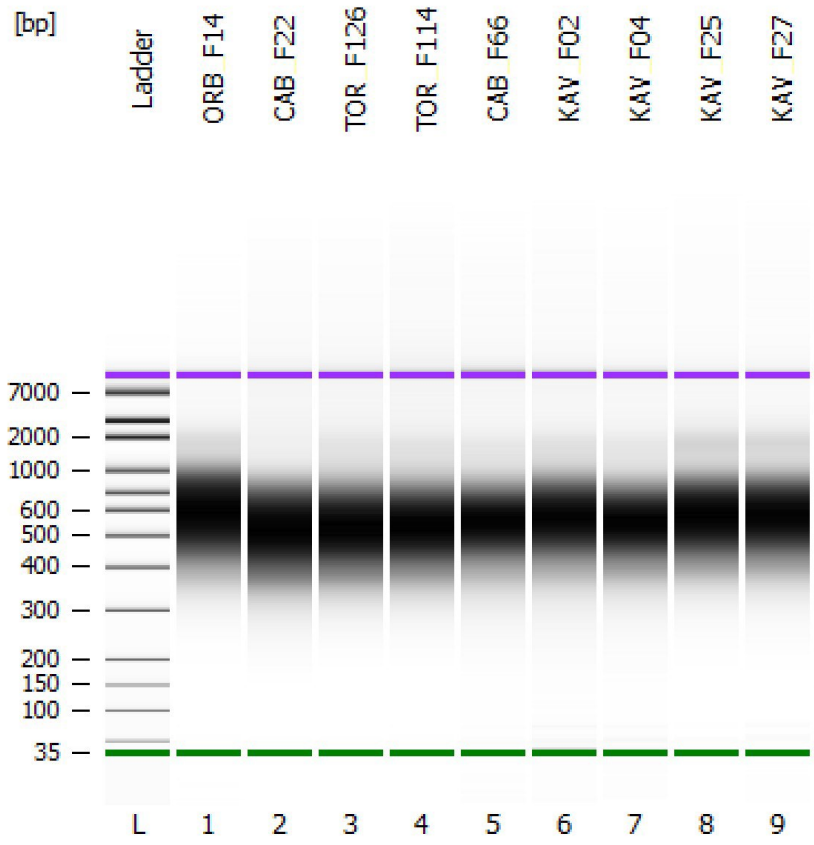


Figura 10. *Esempio di risultati ottenuti dal Bioanalyzer.*

4.3. Qualità dei dati grezzi di sequenziamento

La qualità delle sequenze grezze ottenute dal sequenziamento dei 36 individui di sesso femminile è stata valutata attraverso il software FastQC.

Sono state ottenute una media di 74,9 milioni di *reads* per campione, con un minimo di 58,9 milioni e un massimo di 86,2 milioni di *reads*. Il totale delle *reads* ottenute per tutti i campioni è di 2,39 miliardi.

È stato poi calcolato il *coverage* per ciascun campione. Il *coverage* rappresenta il numero di volte che, in media, ogni posizione di un genoma viene letta durante il sequenziamento ed è stato calcolato attraverso la seguente formula:

$$coverage = \frac{numero\ di\ reads \times lunghezza\ read\ (pb) \times 2}{lunghezza\ genoma\ (pb)}$$

La lunghezza del genoma di *M. cephalus* equivale a 0,8 x 10^9 pb in quanto il genoma è stimato essere 0,8 Gb (Raymond, et al. 2022).

Il *coverage* medio dei campioni analizzati è risultato essere 28X.

Tabella 4. *Per ciascun campione viene riportato il numero di raw reads, cioè le letture ottenute prima del trimming, il numero di paia di basi (raw bp) ottenuto per ciascun campione e il valore del coverage.*

Campione	*Raw reads*	*Raw bp*	*Coverage*
KAV_F02	63.810.517	19.143.155.100	23,93
KAV_F04	64.766.941	19.430.082.300	24,29
KAV_F06	75.565.900	22.669.770.000	28,34
KAV_F07	83.409.095	25.022.728.500	31,28
KAV_F09	84.023.968	25.207.190.400	31,51
KAV_F10	80.927.894	24.278.368.200	30,35
KAV_F11	74.857.141	22.457.142.300	28,07
KAV_F12	78.360.700	23.508.210.000	29,39
KAV_F21	62.500.636	18.750.190.800	23,44
KAV_F22	66.835.037	20.050.511.100	25,06
KAV_F24	65.360.664	19.608.199.200	24,51
KAV_F25	58.912.045	17.673.613.500	22,09
KAV_F27	77.252.797	23.175.839.100	28,97
KAV_F29	75.470.010	22.641.003.000	28,30
KAV_F30	69.578.324	20.873.497.200	26,09
KAV_F31	70.096.525	21.028.957.500	26,29
ORB_F14	80.362.249	24.108.674.700	30,14
ORB_F22	74.335.110	22.300.533.000	27,88
ORB_F23	76.145.320	22.843.596.000	28,55
ORB_F24	80.128.904	24.038.671.200	30,05
ORB_F25	81.913.863	24.574.158.900	30,72
CAB_F16	63.178.382	18.953.514.600	23,69
CAB_F22	85.931.353	25.779.405.900	32,22
CAB_F46	83.982.601	25.194.780.300	31,49
CAB_F66	73.224.181	21.967.254.300	27,46
CAB_F73	85.469.901	25.640.970.300	32,05
TOR_F111	67.151.466	20.145.439.800	25,18
TOR_F114	82.567.293	24.770.187.900	30,96
TOR_F116	71.539.729	21.461.918.700	26,83
TOR_F126	86.218.439	25.865.531.700	32,33
TOR_F128	74.842.524	22.452.757.200	28,07
TOR_F143	81.249.798	24.374.939.400	30,47

I grafici ottenuti da FastQC permettono di definire le sequenze analizzate di qualità elevata, in quanto presentano tutte un *Phread score* superiore a 28.

I principali grafici che vengono creati dal software sono:

- *Per base sequence quality*

  Il grafico permette di visualizzare la qualità, in termini di *Phred score*, delle *reads* dei campioni.

  Per ogni posizione viene disegnato un diagramma di tipo *BoxWhisker*. Questa tipologia di diagramma consiste in una rappresentazione grafica utilizzata per descrivere la distribuzione di un campione ed è costituito dai seguenti elementi:

  - la linea rossa corrisponde alla mediana dei valori relativi alla qualità delle sequenze
  - la linea blu rappresenta la qualità media delle sequenze
  - le barre mostrano l'intervallo assoluto, che copre i valori più bassi (0° quartile) e più alti (4° quartile)

  L'asse x indica la posizione della base nella lettura, mentre l'asse y mostra il punteggio di qualità. Lo sfondo del grafico suddivide l'asse y in tre parti colorate in maniera differente in base alla qualità della chiamata delle basi; il colore verde indica una qualità alta, il giallo media e il rosso bassa.

  In figura 11 viene riportato un esempio di grafico ottenuto dal sequenziamento del campione 'KAV_F25', che risulta ottimale in quanto il punteggio di qualità presenta un minimo di 25 e una media superiore a 34.

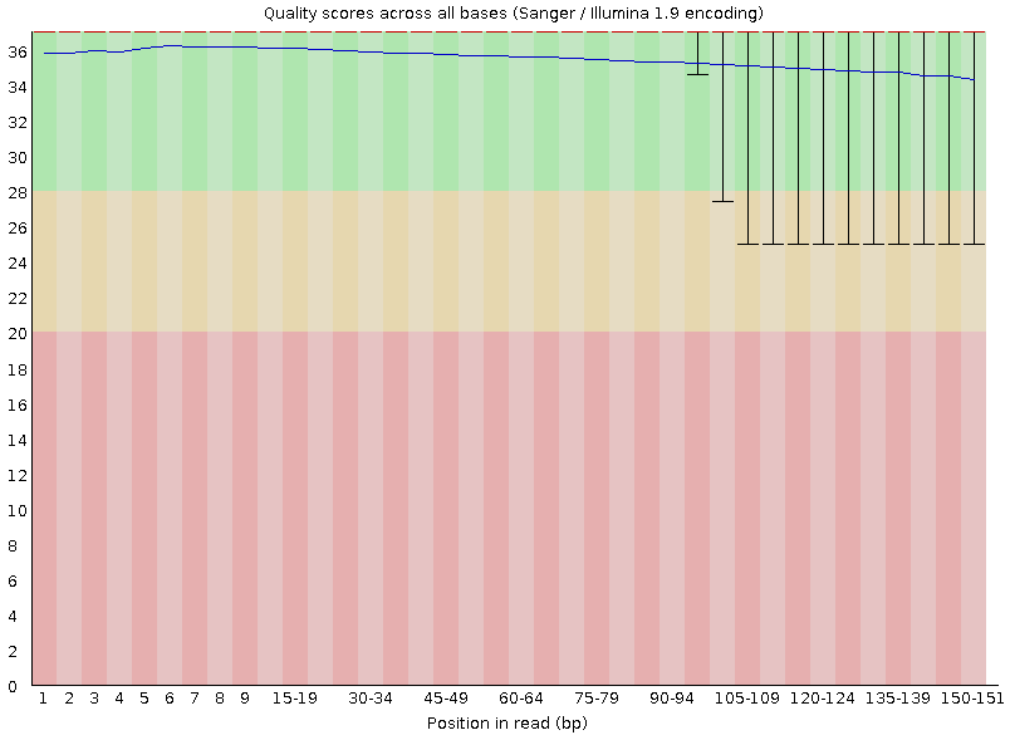


Figura 11. *Esempio di grafico 'Per base sequence quality' (campione 'KAV_F25').*

- *Per sequence quality scores*

  Viene calcolata dal software la qualità media di ciascuna *read* e, attraverso il grafico, si può visualizzarne la distribuzione.

  Dall'esempio riportato in figura 12, si può notare la buona qualità del sequenziamento; il grafico presenta infatti un picco corrispondente ad un *Phread Score* di 36.

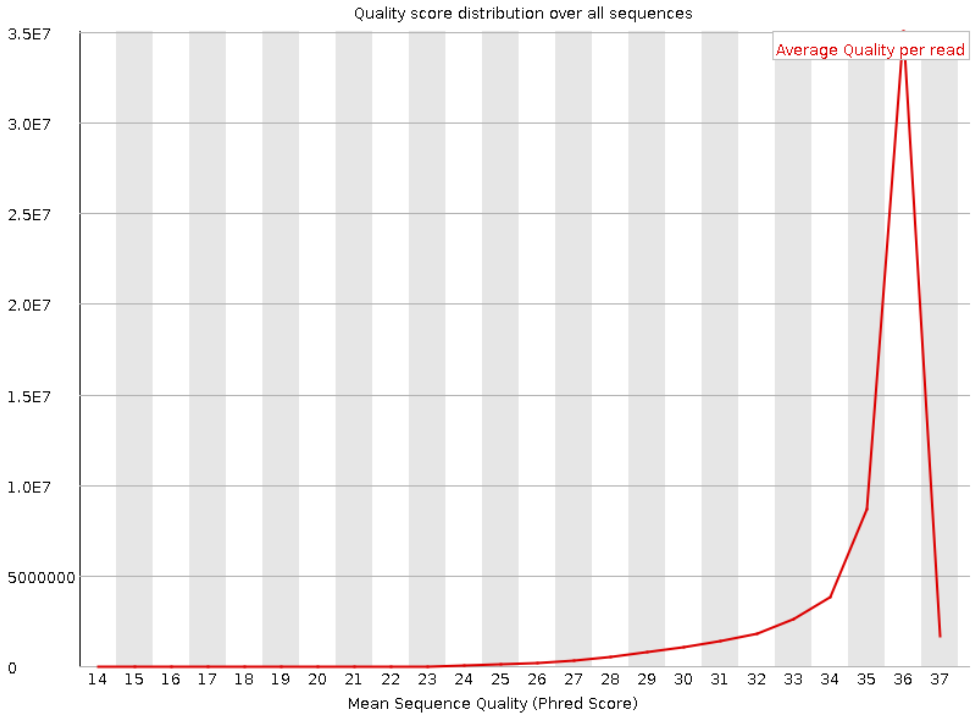


Figura 12. *Esempio di grafico 'Per sequence quality scores' (campione 'KAV_F25').*

- *Per base N content*

  Il grafico permette di valutare la quantità delle basi non identificate, indicate con la lettera N. Per poter definire i dati di buona qualità, il contenuto di N dovrebbe essere <5%.

  Nel campione riportato come esempio (figura 13) non sono presenti basi non identificate.

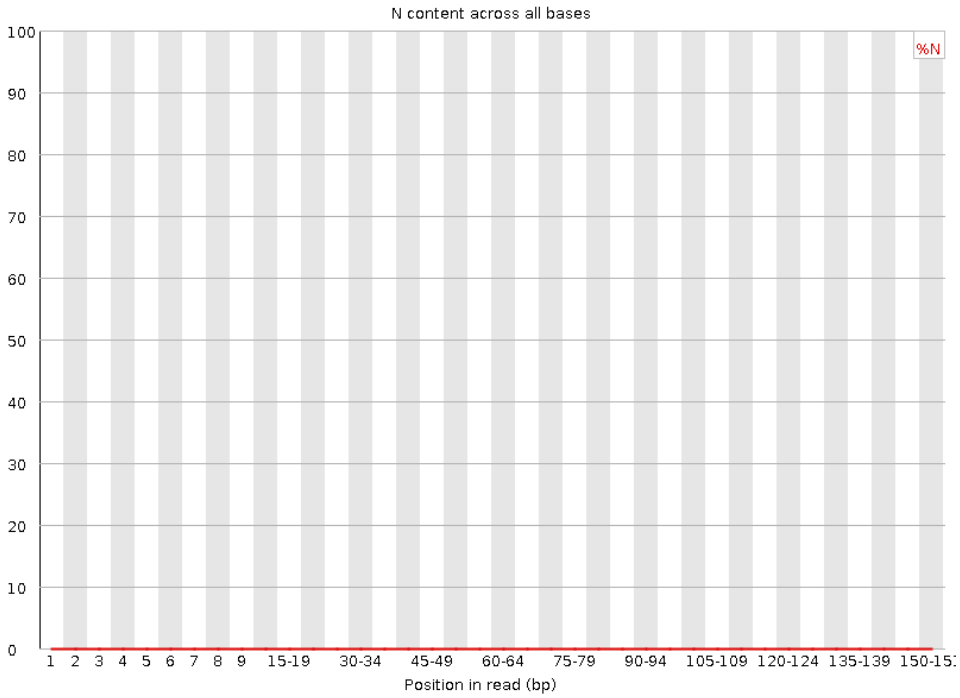


Figura 13. *Esempio di grafico 'Per base N content' (campione 'KAV_F25').*

- *Adapter content*

  Il grafico permette di visualizzare la percentuale di adattatori presenti all'interno delle sequenze per ciascuna posizione. Come si può notare dall'esempio riportato, si riscontra la presenza di una bassa percentuale di adattatori Nextera nelle posizioni finali della sequenza. Questi verranno poi rimossi tramite il processo di *trimming*.

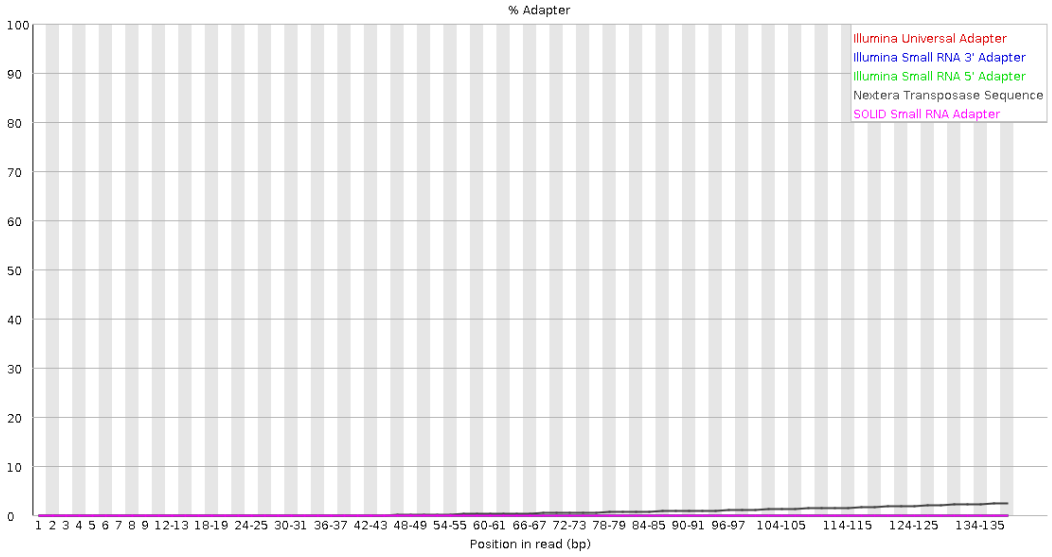


Figura 14. *Esempio di grafico 'Adapter content' (campione 'KAV_F25').*

4.4. Qualità delle sequenze in seguito al *trimming*

Le sequenze grezze sono state sottoposte a *trimming*, con lo scopo di eliminare gli adattatori e le porzioni di bassa qualità. In seguito a questo passaggio, sono state mantenute in media il 98,17% delle *reads*, a conferma della buona qualità del sequenziamento.

Tabella 5. *Per ciascun campione viene riportato il numero di raw reads, cioè le letture ottenute prima del trimming, il numero di trimmed reads, rappresentato dalle sole letture mantenute in seguito all'operazione di trimming e la percentuale delle reads mantenute.*

Campione	*Raw reads*	*Trimmed reads*	% mantenuta
KAV_F02	63.810.517	62.456.218	97,88
KAV_F04	64.766.941	63.563.917	98,14
KAV_F06	75.565.900	74.241.420	98,25
KAV_F07	83.409.095	81.948.840	98,25
KAV_F09	84.023.968	82.451.275	98,13
KAV_F10	80.927.894	79.349.409	98,05
KAV_F11	74.857.141	73.576.737	98,29
KAV_F12	78.360.700	77.014.617	98,28
KAV_F21	62.500.636	61.419.164	98,27
KAV_F22	66.835.037	65.717.426	98,33
KAV_F24	65.360.664	64.189.159	98,21
KAV_F25	58.912.045	57.883.354	98,25
KAV_F27	77.252.797	75.974.361	98,35
KAV_F29	75.470.010	73.975.678	98,02
KAV_F30	69.578.324	68.290.160	98,15
KAV_F31	70.096.525	68.578.293	97,83
ORB_F14	80.362.249	79.069.479	98,39
ORB_F22	74.335.110	73.169.351	98,43
ORB_F23	76.145.320	74.917.899	98,39
ORB_F24	80.128.904	78.752.406	98,28
ORB_F25	81.913.863	80.501.483	98,28
CAB_F16	63.178.382	62.104.188	98,30
CAB_F22	85.931.353	84.373.447	98,19
CAB_F46	83.982.601	82.422.760	98,14
CAB_F66	73.224.181	71.791.210	98,04
CAB_F73	85.469.901	83.944.579	98,22
TOR_F111	67.151.466	65.519.055	97,57
TOR_F114	82.567.293	81.064.676	98,18
TOR_F116	71.539.729	70.155.239	98,06

TOR_F126	86.218.439	84.612.436	98,14
TOR_F128	74.842.524	73.318.254	97,96
TOR_F143	81.249.798	79.848.002	98,27

Sempre tramite il software FastQC è stata valutata la qualità delle sequenze dopo il *trimming*. Dai grafici ottenuti si può notare un aumento della qualità media delle sequenze. Nelle figure 15 e 16 viene riportato un esempio di confronto tra la qualità delle sequenze grezze e la qualità delle sequenze in seguito al processo di *trimming*. Le sequenze di partenza risultano già una buona qualità ma si può comunque notare un miglioramento dopo il *trimming*. Il campione in figura 16 presenta infatti una qualità media superiore a 34 che, al contrario di quanto accade nella sequenza grezza, non diminuisce nelle ultime posizioni.

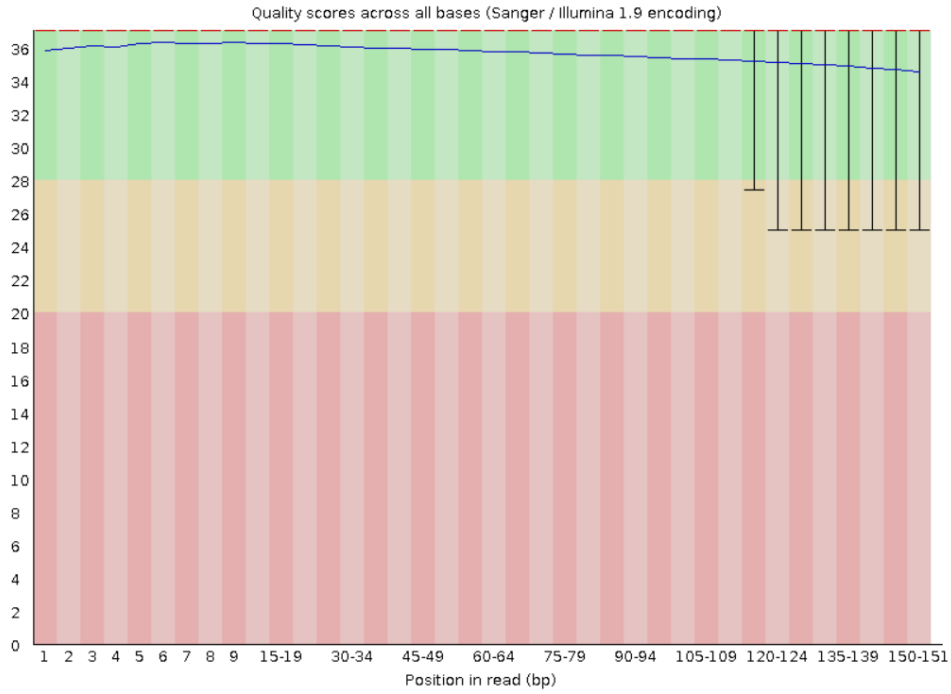


Figura 15. *Esempio di grafico 'Per base sequence quality' di una sequenza grezza (campione 'TOR_F128').*

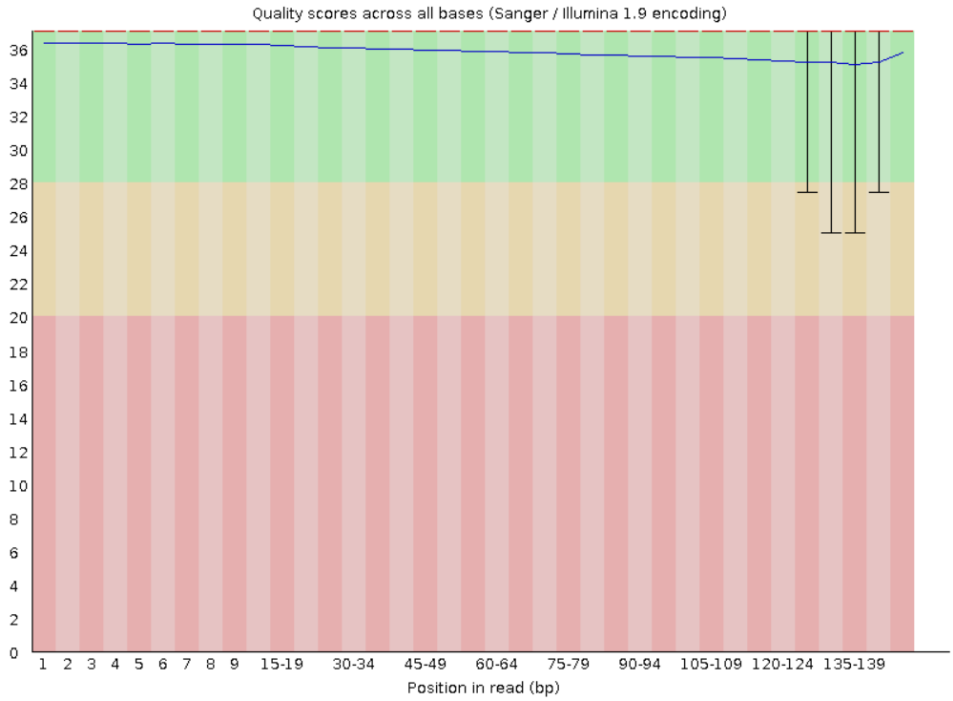


Figura 16. *Esempio di grafico 'Per base sequence quality' ottenuto in seguito all'operazione di trimming (campione 'TOR_F128').*

Il grafico *Sequence length distribution* permette di valutare la distribuzione della lunghezza delle sequenze in seguito al processo di *trimming*. La lunghezza delle sequenze grezze risulta uguale a 151 pb per tutti i campioni, mentre dopo il *trimming* varia a seconda della quantità di basi eliminate (che comprendono gli adattatori e le porzioni di bassa qualità).

In figura 17 si può osservare il grafico ottenuto dal campione 'KAV_F25' in seguito al *trimming*, che presenta un picco a corrispondente alla lunghezza di 144-145 pb.

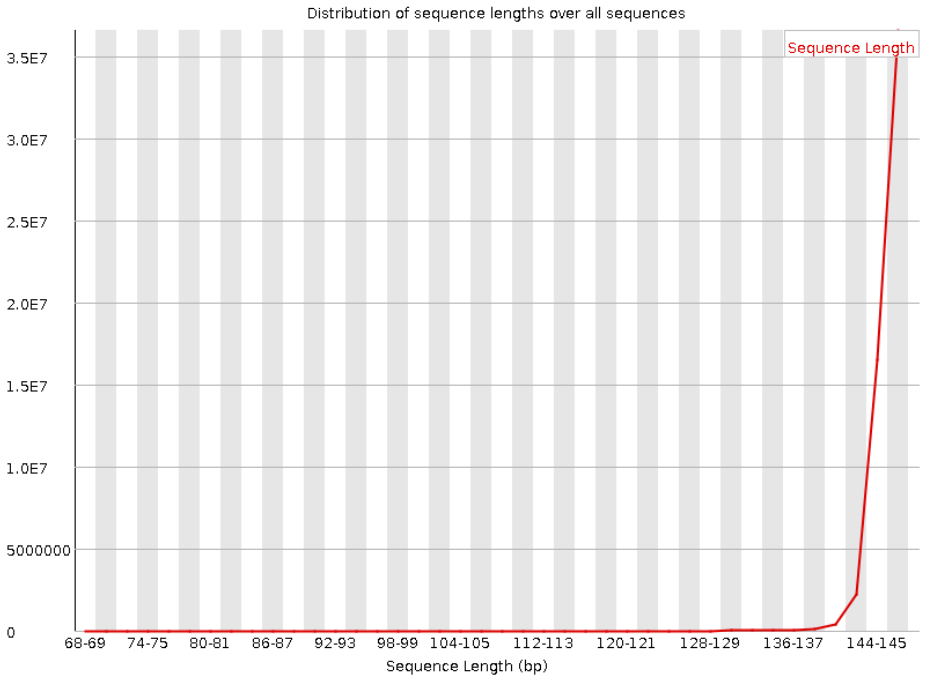


Figura 17. *Esempio di grafico 'Sequence Length Distribution' ottenuto in seguito al processo di trimming (campione 'KAV_F25').*

4.5. *Mapping* delle sequenze sul genoma di riferimento e SNP calling

Le sequenze sono poi state mappate sul genoma di riferimento di *M. cephalus* tramite il software BWA.

Le *reads* correttamente mappate sul genoma sono risultate essere in media 99,74%. Il 97,95% delle *reads* sono state allineate *in pair*, cioè la sequenza *forward* e quella *reverse* sono risultate allineate sulla stessa regione genomica.

Tabella 6. *Per ciascun campione viene riportata la percentuale di reads correttamente mappate sul genoma e la percentuale di reads allineate in pair.*

Campione	% *reads* correttamente mappate sul genoma	% *reads* allineate *in pair*
KAV_F02	99,72	97,90
KAV_F04	99,50	97,59
KAV_F06	99,77	97,96
KAV_F07	99,74	97,81
KAV_F09	99,70	97,88
KAV_F10	99,77	98,01
KAV_F11	99,77	98,00
KAV_F12	99,77	98,04
KAV_F21	99,80	98,00

KAV_F22	99,80	98,11
KAV_F24	99,75	97,92
KAV_F25	99,75	97,89
KAV_F27	99,78	98,03
KAV_F29	99,68	98,01
KAV_F30	99,69	97,92
KAV_F31	99,70	97,92
ORB_F14	99,73	97,82
ORB_F22	99,78	98,01
ORB_F23	99,79	98,05
ORB_F24	99,75	97,79
ORB_F25	99,74	97,96
CAB_F16	99,81	97,94
CAB_F22	99,79	98,03
CAB_F46	99,78	98,04
CAB_F66	99,69	97,90
CAB_F73	99,76	97,93
TOR_F111	99,79	98,03
TOR_F114	99,73	98,00
TOR_F116	99,52	97,69
TOR_F126	99,74	98,00
TOR_F128	99,72	98,13
TOR_F143	99,77	98,01

Si è passati poi alla fase di *SNP calling*, ossia l'identificazione delle varianti nucleotidiche, prendendo in considerazione solamente il contig_2337, contenente il gene *fshr*. Sono stati identificati 25.854 SNP, 5.065 *indels* (tipologia di mutazioni che comprende inserzioni e delezioni di nucleotidi) e 107 siti multiallelici (*indels* e SNP). Spesso l'allineamento in prossimità degli *indel* risulta non preciso e potrebbe portare alla chiamata di SNP non esistenti nella realtà (falsi positivi); sono stati quindi eliminati gli SNP che si trovavano ad una distanza di 5 o meno nucleotidi dagli *indel*. Dopo aver rimosso i siti multiallelici, gli *indel* e gli SNP in loro prossimità, sono risultati analizzabili 23.047 SNP.

4.6. Identificazione di mutazioni legate alla determinazione del sesso

L'ultima fase prevede l'identificazione delle possibili varianti legate alla determinazione del sesso. Questa è stata effettuata analizzando una regione denominata FSHR ± 5 Kb, comprendente il gene *fshr* e la regione genomica compresa tra le 5 Kb a monte e a valle del gene (dalla posizione 369688 alla 391540 del contig_2337).

Due campioni, KAV_M17 e TOR_M196, sulla base del sequenziamento NGS, sono risultati essere maschi conformi, diversamente da quanto era emerso in seguito al sequenziamento effettuato tramite Sanger nel lavoro di Ferraresso et al.. Le successive analisi sono state effettuate escludendo questi due campioni, in quanto probabilmente sono stati erroneamente classificati e avrebbero portato ad una successiva interpretazione non corretta dei dati.

Allo scopo di identificare polimorfismi potenzialmente coinvolti nella determinazione del sesso, è stato deciso di calcolare, per entrambe le popolazioni (Tirreno ed Egeo), l'F_{ST} tra maschi e femmine. Il calcolo dell'indice di fissazione è stato eseguito mantenendo, all'interno delle popolazioni maschili, la frequenza di maschi non conformi osservata in natura, ossia 45% per la popolazione dell'Egeo e 15% per la popolazione del Tirreno.

Tramite i grafici del tipo *Manhattan plot,* eseguiti per ciascuna delle due popolazioni (figura 18), è possibile visualizzare la distribuzione delle mutazioni con F_{ST} più elevato. Per la popolazione dell'Egeo (grafico A) risultano potenzialmente rilevanti i due SNP nelle posizioni 384774 e 384805. La loro presenza non è stata però riscontrata nella popolazione del Tirreno, riducendone quindi l'utilità; ulteriori analisi di approfondimento sarebbero utili per chiarire il possibile ruolo delle due mutazioni nella discriminazione tra i sessi. Il grafico B, che rappresenta la popolazione del Tirreno, permette di evidenziare, tra le varianti con F_{ST} più elevato, la presenza dei tre SNP identificati nei precedenti lavori, nelle posizioni 385101, 385128, 385244. Le tre varianti già note sono state quindi confermate solo in una delle due popolazioni analizzate.

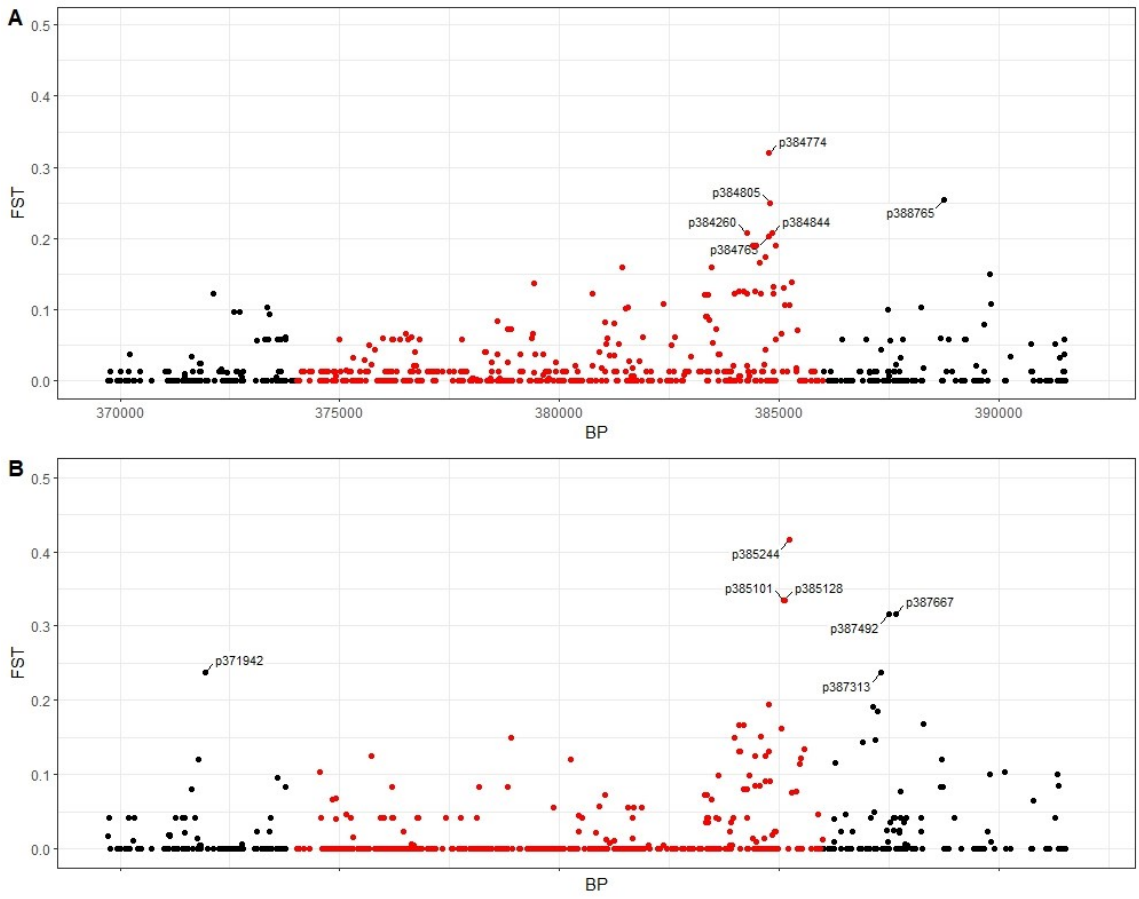


Figura 18. *Manhattan plot per le popolazioni dell'Egeo (A) e del Tirreno (B). Nell'asse delle x viene indicata la posizione (bp), mentre nell'asse y i valori dell' F_{ST}. Ogni puntino corrisponde ad uno SNP; vengono visualizzate le posizioni degli SNP con F_{ST} superiore a 0,2. In rosso sono rappresentate le mutazioni che si trovano all'interno del gene fshr.*

In seguito al calcolo degli F_{ST}, per ciascuna delle due popolazioni (Tirreno ed Egeo), sono stati selezionati i 15 SNP con i valori di F_{ST} più elevati, per valutarne il potenziale ruolo nel differenziare i maschi dalle femmine.

I tre SNP presenti nelle posizioni 385481, 385494 e 387492, identificati nel precedente lavoro come significativi nella distinzione tra maschi conformi e maschi non conformi, sono risultati non rilevanti nel confronto tra individui di sesso maschile e femminile.

Le *heatmap* riportate nelle figure 19 e 20 consentono una visualizzazione dei genotipi di maschi e femmine nelle posizioni selezionate. Per la popolazione dell'Egeo (figura 19) la mutazione con F_{ST} più elevato è alla posizione 384774; non è sufficiente però a differenziare i due sessi. Nella *heatmap* relativa agli individui del Tirreno (figura 20) si può notare una evidente differenza tra maschi (wt/mut) e femmine (wt/wt) nelle posizioni 385244, 385101 e 385128.

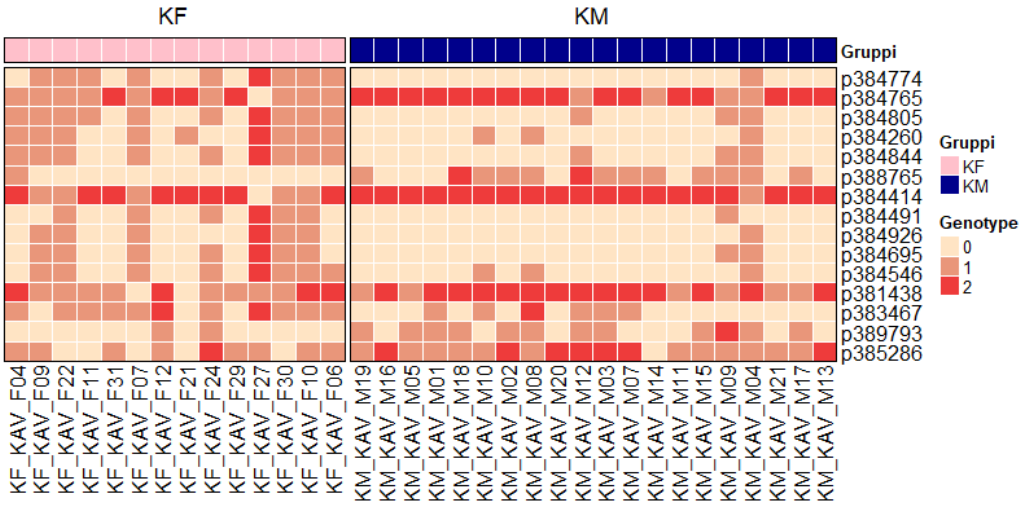


Figura 19. *Heatmap relativa agli individui appartenenti alla popolazione dell'Egeo. (Genotipo: 0=wt/wt, 1=wt/mut, 2=mut/mut)*

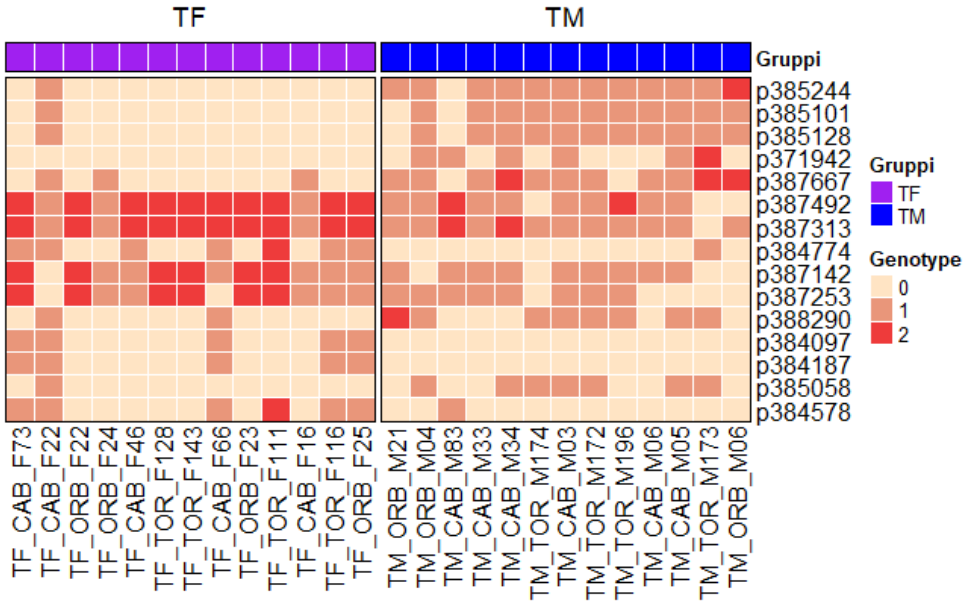


Figura 20. *Heatmap relativa agli individui appartenenti alla popolazione del Tirreno. (Genotipo: 0=wt/wt, 1=wt/mut, 2=mut/mut)*

La figura 21 rappresenta la *heatmap* complessiva che comprende entrambe le popolazioni e le 30 posizioni più rilevanti, derivanti dall'unione delle 15 mutazioni con valore di F_{ST} più elevato di ciascuna popolazione. Tra queste solo la variante in posizione 384774 è risultata essere presente in entrambe le popolazioni, ma non è comunque particolarmente significativa in quanto non appare rilevante per la popolazione del Tirreno.

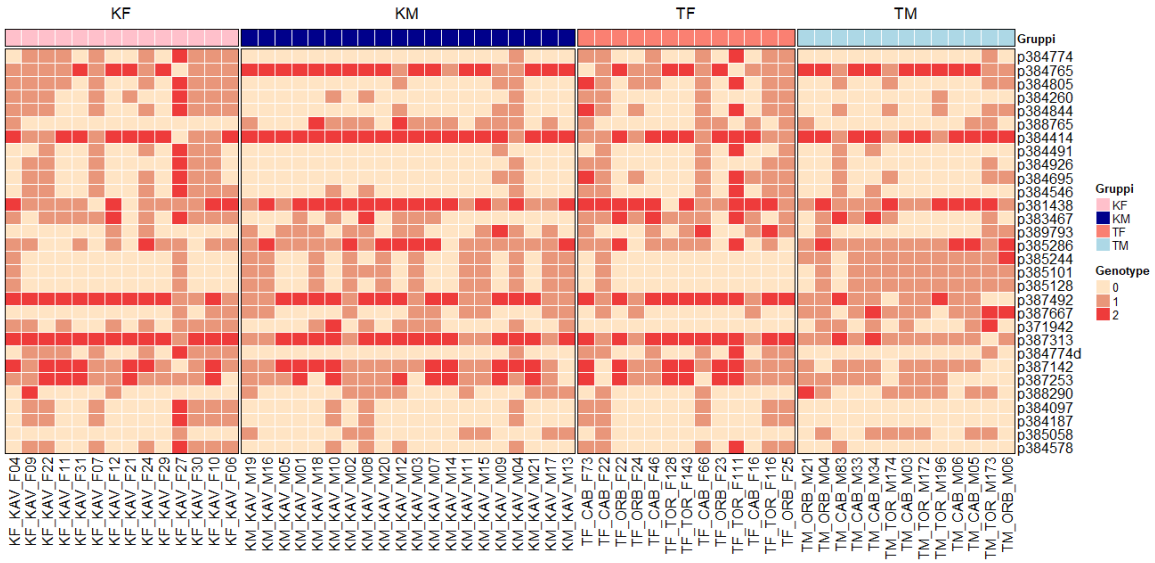


Figura 21. *Heatmap relativa alle popolazioni di Tirreno ed Egeo. Vengono riportati i genotipi nelle 30 posizioni selezionate. (Genotipo: 0=wt/wt, 1=wt/mut, 2=mut/mut)*

Nella *heatmap* riportata in figura 22 viene evidenziata, per entrambe le popolazioni, il genotipo per lo SNP 384774 con quello degli SNP 384101, 384128 e 384244, originariamente proposti come varianti determinanti il sesso nel cefalo. La mutazione in posizione 384774 risulta essere presente in omozigosi per uno dei due alleli nella quasi totalità degli animali di sesso fenotipico maschile, mentre nelle femmine si osserva frequentemente il genotipo eterozigote e in due casi il genotipo omozigote per l'allele alternativo. Utilizzando l'insieme dei genotipi per queste quattro varianti è possibile classificare tutti i maschi con la regola che abbiano genotipo 0 per 384744 e/o genotipo 1 o 2 per uno degli SNP 384101, 384128 e 384244. Le femmine possono essere classificate tutte correttamente, ad eccezione di KAV_F04, se hanno genotipo 1 o 2 per 384744 e/o genotipo 0 per uno o più degli SNP 384101, 384128 e 384244.

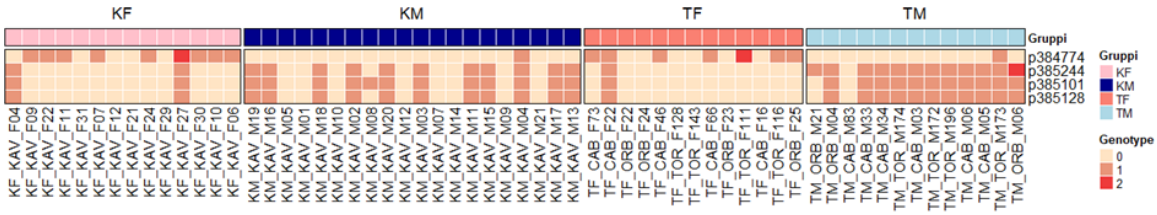


Figura 22. *Heatmap relativa alle popolazioni di Tirreno ed Egeo. Vengono riportati i genotipi nelle 4 posizioni 384774, 384101, 384128 e 384244. (Genotipo: 0=wt/wt, 1=wt/mut, 2=mut/mut)*

5. Discussione

Il cefalo risulta essere una tra le specie più promettenti per l'acquacoltura, in particolare per l'importanza delle gonadi femminili che sono la base per la produzione della bottarga, un prodotto alimentare molto rinomato. La bottarga di muggine si ottiene mediante un processo di salagione ed essiccazione delle ovaie e viene venduta ad un prezzo che oscilla tra i 150 e i 300 €/Kg; visto il suo elevato valore economico, riuscire ad avere allevamenti di cefalo che comprendano principalmente individui di sesso femminile, porterebbe ad enormi vantaggi. Tuttavia la maturazione sessuale in *M. cephalus* viene raggiunta a circa due anni di età; l'identificazione di una strategia che permetta di determinare con elevata probabilità il sesso degli individui, permetterà di allevare solo le femmine fin dai primissimi stadi di sviluppo. Ad oggi i meccanismi che regolano la determinazione del sesso in *M. cephalus* risultano ancora poco chiari.

Dor et al. (2016), analizzando marcatori genetici microsatelliti sulla progenie di *M. cephalus* in due famiglie indipendenti hanno dimostrato che gli alleli legati al sesso maschile vengono trasmessi dal padre. Sono stati inoltre individuati cinque geni potenzialmente associati alla determinazione del sesso: *gth-ri* (anche conosciuto come *fshr*), *foxi1*, *dhx32*, *bub3* e *dock1* (L. Dor, et al. 2016). In uno studio successivo, i geni *bccip*, *dhx32a*, *dock1* e *fshr* (*gth-ri*) sono stati ipotizzati come regolatori della determinazione del sesso (L. Dor, et al. 2020).

Il gene *fshr* codifica per un recettore transmembrana, appartenente alla famiglia dei recettori accoppiati alle proteine G; il suo ligando è l'ormone follicolo stimolante (FSH). L'azione di FSH è mediata dal legame con il suo recettore specifico, localizzato sulla membrana cellulare delle cellule del Sertoli nel testicolo e delle cellule della granulosa nei follicoli ovarici.

In un lavoro condotto su zebrafish (*Danio rerio*) sono stati prodotti mutanti *knock-out* per il gene *fshr*. Gli individui di sesso femminile hanno mostrato un fenomeno di *sex reversal*, mantenendo un'apparente fertilità, mentre i maschi non hanno subito cambiamenti rilevanti. I risultati ottenuti dallo studio suggeriscono un ruolo chiave del gene *fshr* in zebrafish e nei pesci teleostei in generale (Zhang , et al. 2015).

Anche in altre specie di pesci, ad esempio nella sogliola atlantica (*Solea senegalensis*), è stato identificato *fshr* come gene determinante il sesso. In questo caso è stato ipotizzato che la variante legata al cromosoma Y (*fshry*), possa ridurre il segnale dell'ormone follicolo stimolante, causando una sovraregolazione dell'ormone

antimülleriano e quindi inducendo la gonade indifferenziata verso la maturazione in testicolo (de la Herrán, et al. 2022).

Il gene *fshr* è stato confermato essere legato alla determinazione del sesso nel cefalo anche in un ulteriore studio (Ferraresso, et al. 2021), nel quale l'analisi tramite Pool-Seq di individui di *M. cephalus* appartenenti a diverse popolazioni ha permesso di individuare tre mutazioni legate al sesso, presenti all'interno del gene *fshr,* denominate MuCe179, MuCe206 e MuCe322. Le mutazioni in posizione 179 e 206 sono mutazioni missenso, mentre nella posizione 322 è presente una mutazione sinonima. Le due mutazioni missenso risultano presenti in eterozigosi nel fenotipo maschile. Mostrano però una penetranza incompleta in un determinato numero di maschi che risultano avere un genotipo wt/wt, associato invece al fenotipo femminile e per questo vengono denominati maschi "non conformi".

In un lavoro successivo, mediante la tecnologia Illumina è stato sequenziato l'intero genoma di maschi conformi e maschi non conformi di due popolazioni, permettendo di assegnare in modo preciso il genotipo a ciascun animale. Le varianti MuCe179, MuCe206 e MuCe322 (nelle posizioni 385101, 385128 e 385244 della versione aggiornata del genoma di *M. cephalus* utilizzata nell'analisi) sono state confermate; inoltre sono state identificate tre nuove mutazioni, due nel gene *fshr* e una nell'immediata regione a valle, che presentano una frequenza allelica significativamente diversa tra maschi conformi e maschi non conformi. Gli SNP in posizione 385481 e 385494 risultano essere presenti nel 95% dei maschi conformi e nel 16% dei maschi non conformi analizzati. Lo SNP in posizione 387492 mostra invece una differente frequenza allelica tra i maschi non conformi appartenenti alle due differenti popolazioni e risulta quindi non utile nella differenziazione tra i sessi.

Il presente lavoro di tesi si è concentrato nell'analisi degli individui di sesso femminile appartenenti alle due differenti popolazioni, per poterli confrontare con i maschi già precedentemente analizzati ed avere una visione completa dei possibili polimorfismi alla base della determinazione del sesso in *M. cephalus.*

Tramite l'analisi degli individui appartenenti alla popolazione dell'Egeo, è stata riscontrata la presenza di due SNP con un F_{ST} elevato nelle posizioni 384774 e 384805. Tuttavia all'interno della popolazione del Tirreno la mutazione nella posizione 384805 presenta un valore di F_{ST} basso, che la rende non informativa nell'ambito della differenziazione tra i sessi. I tre SNP nelle posizioni 385101, 385128 e 385244 identificati nel precedente lavoro di Ferraresso et al. risultano avere un F_{ST} elevato nella popolazione del Tirreno, consentendo di confermare la loro rilevanza nella distinzione

tra maschi e femmine per la popolazione analizzata. All'interno della porzione di genoma presa in considerazione (FSHR ± 5 Kb) non sono state trovate singole mutazioni significative per entrambe le popolazioni che permettano di differenziare in maniera chiara gli individui di sesso femminile da quelli di sesso maschile.

La mutazione in posizione 384774 tuttavia risulta essere presente in omozigosi per uno dei due alleli in quasi tutti i maschi appartenenti ad entrambe le popolazioni. Questo SNP, confrontato con gli SNP 384101, 384128 e 38424, potrebbe aiutare nell'identificazione degli individui di sesso maschile. Come già descritto, la presenza di una combinazione di genotipi (0 per 384744 e/o 1/2 per 384101/384128/38424) potrebbe essere utilizzata per identificare i maschi, mentre la combinazione 1/2 per 384744 e/o 0 per 384101/384128/38424 classifica tutte le femmine con una singola eccezione. Questo risultato e il conseguente saggio genetico di rivelazione del sesso deve tuttavia essere convalidato in un numero più ampio di individui e di popolazioni. Anche se confermato, questo tipo di determinazione del sesso richiederebbe ulteriori studi per comprenderne i meccanismi molecolari di base. Una prima ipotesi potrebbe prevedere che la mutazione 384744, localizzata a livello intronico, regoli l'espressione del gene (l'omozigote 0 potrebbe avere un'espressione ridotta), mentre le mutazioni 384101/384128/38424 in eterozigosi potrebbero, come già ipotizzato, avere un effetto dominante-negativo sulla funzione della proteina, sommandosi con effetto additivo.

Considerati i risultati ottenuti da questo e dai precedenti studi, non si può escludere che il cefalo sia caratterizzato da una determinazione del sesso poligenica (PSD) e/o influenzata da fattori ambientali (ESD). Per "determinazione del sesso poligenica" si intende la presenza di diversi loci distribuiti nel genoma che influenzano il sesso di ciascun individuo, come avviene ad esempio nella spigola (*Dicentrarchus labrax*) (Saillant, et al. 2002), un pesce teleosteo appartenente alla famiglia Moronidae. Questa è una specie gonocorica con un sistema di determinazione del sesso poligenico influenzato da fattori ambientali, principalmente la temperatura (Vandeputte e Piferrer 2019).

È probabile che la determinazione genetica del sesso in *M. cephalus* venga influenzata anche da fattori ambientali. Questo accade anche in molte altre specie, tra cui il pesce denominato Chinese tongue sole (*C. semilaevis*), nel quale i fattori ambientali attivano meccanismi di regolazione epigenetica (Shao, et al. 2014).

Nonostante sia chiaro che il gene *fshr* abbia un ruolo fondamentale nella determinazione del sesso nel cefalo, non si può escludere la presenza di altri geni coinvolti. Appare quindi evidente la necessità di ulteriori analisi per verificare se esiste una metodica che

consenta il sessaggio genetico della specie in esame. Analisi aggiuntive che prendano in considerazione l'intero genoma potrebbero portare all'identificazione di altri geni che contribuiscono alla differenziazione tra maschi e femmine. Sarebbe inoltre opportuno estendere l'analisi ad un maggior numero di individui provenienti da differenti popolazioni e svolgere ulteriori studi sul possibile effetto di fattori ambientali.

6. Conclusione

Nel presente lavoro di tesi è stato sequenziato l'intero genoma di 32 campioni di cefalo di sesso femminile, 16 per popolazione, per confrontarli con i dati di individui maschi appartenenti alle stesse popolazioni, già sequenziati in precedenza. Lo scopo principale era quello di investigare l'esistenza di ulteriori loci capaci di influenzare la determinazione del sesso in *M. cephalus*. Considerando diversi studi presenti in letteratura, è stata definita una probabile funzione del gene *fshr* nel complesso meccanismo di determinazione del sesso in *M. cephalus*. Le prime analisi si sono quindi focalizzate nel gene *fshr* e nella regione genomica immediatamente contigua (5Kb a monte e a valle del gene), denominata FSHR ± 5Kb.

Le analisi effettuate hanno evidenziato due scenari diversi a seconda dell'origine geografica degli animali analizzati. Nella popolazione del Tirreno è stata confermata la rilevanza degli SNP presenti nelle posizioni 385101, 385128 e 385244 e la loro utilità nella discriminazione tra maschi e femmine con una accuratezza del 92%. Queste varianti non sono invece risultate essere tra le più distintive tra i due sessi nella popolazione dell'Egeo. In quest'ultima popolazione, le analisi effettuate hanno permesso di individuare due mutazioni nelle posizioni 384774 e 384805 potenzialmente rilevanti per la diversificazione tra maschi e femmine avendo rispettivamente un'accuratezza del 83% e del 76%. Nonostante queste ultime due mutazioni permettano di incrementare notevolmente l'accuratezza nella discriminazione tra maschi e femmine nella popolazione dell'Egeo (portandola dal 45% al 83%), non possono essere considerate come il locus determinante la differenziazione sessuale in *M. cephalus*. È stata inoltre osservata una possibile correlazione tra lo SNP in posizione 384774 e gli SNP 384101, 384128 e 38424 per gli individui di sesso maschile di entrambe le popolazioni. Queste mutazioni, tramite un effetto additivo, potrebbero influenzare la funzionalità della proteina.

Nonostante il presente studio non abbia permesso di arrivare ad una completa comprensione dei complessi meccanismi che regolano la determinazione del sesso in *M. cephalus*, le analisi effettuate hanno comunque permesso di ampliare le conoscenze riguardanti la determinazione genetica del sesso nella specie analizzata e proporre una combinazione genotipica per il sessaggio degli individui immaturi.

L'obiettivo di individuare un metodo che consenta il sessaggio precoce dei cefali per ottenere popolazioni contenenti in gran parte o totalmente individui di sesso femminile, potrà essere raggiunto tramite analisi più approfondite a livello dell'intero genoma.

Sarebbe opportuno inoltre estendere i successivi studi ad un numero più elevato di individui provenienti da differenti popolazioni.

7. Bibliografia

Capel, Blanche. «Vertebrate sex determination: evolutionary plasticity of a fundamental switch.» *Nature Reviews Genetics*, 2017: 675–689.

Chen, Ji, Zuoyan Zhu, e Wei Hu. «Progress in research on fish sex determining genes.» *KeAi Water Biology and Security*, 2022.

Cortez, Diego, et al. «Origins and functional evolution of Y chromosomes across mammals.» *Nature*, 2014: 488–493.

Crosetti, Donatella. «Current state of grey mullet fisheries and culture.» In *Biology, Ecology and Culture of Grey Mullet (Mugilidae)*, di Donatella Crosetti e Stephen J. M. Blaber, 398–450. CRC Press, 2016.

Danecek, Petr, et al. «The variant call format and VCFtools.» *Bioinformatics Applications Note*, 2011: 2156–2158.

Danecek, Petr, et al. «Twelve years of SAMtools and BCFtools.» *GigaScience*, 2021: 1-4.

de la Herrán, Roberto, et al. «A chromosome-level genome assembly enables the identification of the follicule stimulating hormone receptor as the master sex-determining gene in the flatfish Solea senegalensis.» *Molecular Ecology Resources*, 2022: 886–904.

Dor, L., et al. «Identification of the sex-determining region in flathead grey mullet (Mugil cephalus).» *Animal Genetics* , 2016: 698–707.

Dor, Lior, et al. «Preferential Mapping of Sex-Biased Differentially-Expressed Genes of Larvae to the Sex-Determining Region of Flathead Grey Mullet (Mugil cephalus).» *Frontiers in Genetics*, 2020: 839.

Ferraresso, Serena, et al. «fshr: a fish sex-determining locus shows variable incomplete penetrance across flathead grey mullet populations.» *iScience*, 2021.

Ge, Chutian, et al. «The histone demethylase KDM6B regulates temperature-dependent sex determination in a turtle species.» *Science* , 2018: 645–648.

Giani, Alice Maria, Guido Roberto Gallo, Luca Gianfranceschi, e Giulio Formenti. «Long walk to genomics: History and current approaches to genome sequencing and assembly.» *Computational and Structural Biotechnology Journal*, 2020: 9-19.

Hung, Chih-Ming, e Daigee Shaw. «The Impact of Upstream Catch and Global Warming on the Grey Mullet Fishery in Taiwan: A Non-cooperative Game Analysis.» *Marine Resource Economics*, 2006: 285–300.

Kang, Yuna, Chon-Sik Kang, e Changsoo Kim. «History of Nucleotide Sequencing Technologies: Advances in Exploring Nucleotide Sequences from Mendel to the 21st Century.» *Horticultural Science and Technology*, 2019: 549-558.

Koopman, Peter, John Gubbay, Nigel Vivian, Peter Goodfellow, e Robin Lovell-Badge. «Male development of chromosomally female mice transgenic for Sry.» *Nature*, 1991: 117-121.

Li, Heng. «A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.» *Bioinformatics*, 2011: 2987–2993.

Li, Heng, e Richard Durbin. «Fast and accurate long-read alignment with Burrows–Wheeler transform.» *Bioinformatics Original Paper*, 2010: 589–595.

Li, Jianzhen, e Christopher H.K. Cheng. «Evolution of gonadotropin signaling on gonad development: Insights from gene knockout studies in zebrafish.» *Biology of Reproduction*, 2018: 686 - 694.

Li, Xi-Yin, e Jian-Fang Gui. «Diverse and variable sex determination mechanisms in vertebrates.» *SCIENCE CHINA Life Sciences*, 2018: 1503-1514.

Raymond, Jani Angel J., et al. «Comparative genome size estimation of different life stages of grey mullet, Mugil cephalus Linnaeus, 1758 by flow cytometry.» *Aquaculture Research*, 2022: 1151-1158.

Rens, Willem, et al. «The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z.» *Genome Biology*, 2007: R243.

Rens, Willem, Frank Grützner, Patricia C. M. O'Brien, Helen Fairclough, Jennifer A. M. Graves, e Malcolm A. Ferguson-Smith. «Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution.» *PNAS*, 2004: 16257–16261.

Ribas , Laia, Woei Chang Liew, Noèlia Díaz, Rajini Sreenivasan, László Orbán, e Francesc Piferrer. «Heat-induced masculinization in domesticated zebrafish is family-specific & yields a set of different gonadal transcriptomes.» *Proceedings of the National Academy of Sciences of the United States of America*, 2017: E941 - E950.

Saillant, Eric , Alexis Fostier, Pierrick Haffray, Bruno Menu, Jacques Thimonier, e Béatrice Chatain. «Temperature effects and genotype-temperature interactions on sex determination in the European sea bass (Dicentrarchus labrax L.).» *Journal of Experimental Zoology*, 2002: 494-505.

Sanger, F., S. Nicklen, e A. R. Coulson. «DNA sequencing with chain-terminating inhibitors.» *Proceedings of the National Academy of Sciences of the United States of America*, 1977: 5463-5467.

Sato, Youichi, Toshikatsu Shinka, Kozue Sakamoto, Ashraf A. Ewis, e Yutaka Nakahori. «The male-determining gene SRY is a hybrid of DGCR8 and SOX3, and is regulated by the transcription factor CP2.» *Mol Cell Biochem*, 2010: 267-275.

Schartl, Manfred, et al. «Sox5 is involved in germ-cell regulation and sex determination in medaka following co-option of nested transposable elements.» *BMC Biology*, 2018: 16.

Shao, Changwei, et al. «Epigenetic modification and inheritance in sexual reversal of fish.» *Genome Research*, 2014: 604-615.

Smith, Craig A., et al. «The avian Z-linked gene DMRT1 is required for male sex determination in the chicken.» *Nature*, 2009: 267–271.

Sutou, S., Y. Mitsui, e K. Tsuchiya. «Sex determination without the Y Chromosome in two Japanese rodents Tokudaia osimensis osimensis and Tokudaia osimensis spp.» *Mammalian Genome*, 2001: 17-21.

Sutton, Edwina, et al. «Identification of SOX3 as an XX male sex reversal gene in mice and humans.» *The Journal of Clinical Investigation*, 2011: 328-341.

Usman, Muhammad, et al. «Introduction to world production of fish roe and preocessing.» In *Fish Roe: Biochemistry, Products, and Safety*, di Alaa El-Din A. Bekhit, 430. Alaa El-Din A. Bekhit, 2022.

Vandeputte, Marc, e Francesc Piferrer. «Genetic and Environmental Components of Sex Determination in the European Sea Bass.» In *Sex Control in Aquaculture*, di Hanping Wang, Francesc Piferrer , Songlin Chen e Zhi-Gang Shen , 307-325. 2019.

Wessels, Stephan, et al. «Allelic Variant in the Anti-Mu¨ llerian Hormone Gene Leads to Autosomal and Temperature-Dependent Sex Reversal in a Selected Nile Tilapia Line.» *PLoS One 9*, 2014: e104795.

Zhang , Zhiwei , Shuk-Wa Lau, Lingling Zhang, e Wei Ge. «Disruption of Zebrafish Follicle-Stimulating Hormone Receptor (fshr) But Not Luteinizing Hormone Receptor (lhcgr) Gene by TALEN Leads to Failed Follicle Activation in Females Followed by Sexual Reversal to Males.» *Endocrinology*, 2015: 3747–3762.

8. Sitografia

https://www.agraria.org/pesci/muggine.htm

https://www.fao.org/

https://www.illumina.com/

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

https://broadinstitute.github.io/picard/

https://posit.co/products/open-source/rstudio/