UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di Medicina Animale, Produzioni e Salute
Dipartimento di Biomedicina Comparata e Alimentazione

Corso di laurea magistrale a ciclo unico in Medicina Veterinaria

Ricerca di loci legati alla determinazione del sesso in *Mugil cephalus* tramite whole genome sequencing

Relatore
Prof. Tomaso Patarnello

Co-relatore
Dr.ssa Serena Ferraresso

Laureanda
Silvia Gobbi
Matricola n.
1163088

ANNO ACCADEMICO 2021/2022

Riassunto

Nei pesci teleostei, i meccanismi di determinazione del sesso sono estremamente diversi, spaziando dalla presenza di cromosomi sessuali all'ermafroditismo simultaneo. Anche nei casi di determinazione genetica del sesso (GSD; *Genetic Sex Determination*), i meccanismi possono differire anche tra specie filogeneticamente vicine. Inoltre, le condizioni ambientali possono contribuire alla determinazione del sesso diventandone talvolta i fattori predominanti (ESD; *Environmental Sex Determination*).

Il cefalo (*Mugil cephalus*) è un pesce osseo di notevole interesse sia per la pesca che per l'acquacultura, soprattutto in paesi come il Giappone, Taiwan, Korea e in alcune aree del Mar Mediterraneo. In queste ultime aree il cefalo assume un valore particolarmente rilevante grazie alla produzione della bottarga, una specialità culinaria molto pregiata e dall'alto valore economico costituita dalla gonade femminile pressata, salata ed essiccata.

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.

Il cefalo è una specie a sessi separati (gonocorica) con determinazione genetica del sesso; tuttavia, diversi studi ne hanno riportato l'assenza di cromosomi sessuali. In un recente studio basato sull'analisi dell'intero genoma, sono state identificate, su popolazioni di diversa origine geografica, due mutazioni missenso (mutazioni puntiformi o SNP, che determinano un cambiamento aminoacidico) localizzate in *cis* nel gene codificante per il recettore dell'ormone follicolo-stimolante (*fshr*). Queste mutazioni, chiamate M1, sono significativamente associate al sesso fenotipico. Infatti M1 è presente in eterozigosi (wt/m1) nella gran parte degli individui con fenotipo maschile. Un certo numero di maschi, definiti "non-conformi", hanno però mostrato un genotipo wt/wt, normalmente associato al fenotipo femminile. La frequenza di maschi non-conformi si è mostrata relativamente bassa (6-18%) nelle popolazioni italiane analizzate, mentre ha raggiunto una frequenza prossima al 50% in una popolazione dell'Egeo. In questo quadro, il gene *fshr* pur avendo certamente un ruolo nella determinazione del sesso di *M. cephalus* (soprattutto in alcune popolazioni) non ne è l'unico "determinante". Ciò suggerisce il possibile coinvolgimento anche di altri loci nella determinazione del sesso in questa specie. Il gene *fshr* viene perciò definito "a penetranza incompleta" relativamente al suo ruolo nella capacità di influenzare il sesso nel cefalo.

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, eventualmente in altri loci genici, capaci di influenzare la determinazione del sesso in *M. cephalus*.

Per raggiungere tale scopo, il genoma di maschi conformi e non-conformi di diverse popolazioni è stato sequenziato a medio coverage (25X) e le varianti genetiche identificate sono state confrontate tra i due gruppi.

Abstract

In teleost fish, mechanisms of sex determination are extremely different, ranging from presence of sex chromosomes to simultaneous hermaphroditism. Even in cases of Genetic Sex Determination (GSD), mechanisms can also differ between phylogenetically close species. Furthermore, environmental conditions can contribute to the determination of sex, sometimes becoming predominant factors (ESD; Environmental Sex Determination).

The flathead grey mullet (*Mugil cephalus*) is a bony fish of considerable interest both for fishing and aquaculture, especially in countries such as Japan, Taiwan, Korea and in some areas of the Mediterranean Sea. In the latter areas mullets take on a significant value thanks to the production of bottarga, a highly prized culinary specialty with a high economic value consisting of pressed, salted and dried female gonad.

The ability to control sex ratio, with an almost exclusively female production, and / or the identification of genetic markers capable of prematurely sexing individuals would lead enormous benefits to the flathead grey mullet industry.

The flathead grey mullet is a species with separate sexes (gonochoric) with genetic sex determination, however several studies have reported the absence of sex chromosomes. In a recent study based on the analysis of whole genome, two missense mutations (point mutations or SNPs, which result in an amino acid exchange), *cis*-located in the gene coding for the follicle-stimulating hormone receptor (*fshr*) were identified on populations of different geographical origin. These mutations, named M1, were found to be significantly associated with phenotypic sex. M1 is heterozygous (wt / m1) in majority of male phenotype. A certain number of males, defined as “non-conformi”, exhibit a wt / wt genotype, associated instead with the female phenotype. The frequency of wt /wt males was relatively low (6-18%) in the italian populations analysed, while it reached a frequency close to 50% in a population of the Aegean Sea. From this point of view, *fshr* gene, while certainly has a sex determining role in *M. cephalus* (especially in some populations), is not the only "determinant". *fshr* gene is therefore defined as "incomplete penetrance" in relation to its role in the ability to influence sex in flathead grey mullet.

A different genetic heritage, as well as different environmental conditions, can play an important role in determining the sex in *M. cephalus*, modulating the action of *fshr*. In this context, the aim of this thesis was to evaluate the existence of additional genetic loci involved in the *M. cephalus* sex determination.

To achieve this, the genome of wt / wt and wt / m1 males from different populations was sequenced at medium coverage (25X) and the identified genetic variants were compared between two groups.

Indice

1. Introduzione

1.1. Meccanismi di determinazione del sesso nei vertebrati

La maggior parte degli organismi eucarioti si riproduce sessualmente e i fenotipi riproduttivi maschili e femminili sono ampiamente conservati attraverso i differenti taxa (Pennell, et al., 2018). All'interno dei vertebrati i sistemi di riproduzione sono molto differenti, spaziando da organismi con sessi separati (gonocorismo ed ermafroditismo sequenziale) all'ermafroditismo simultaneo (Eppley e Jesson, 2008).

La maggior parte dei vertebrati presenti in natura manifesta una sessualità con sessi separati (gonocorismo) (Herpin e Schartl, 2015), la cui differenziazione può avvenire attraverso una determinazione genetica del sesso (GSD; *Genetic Sex Determination*), una determinazione ambientale del sesso (ESD; *Environmental Sex Determination*), o una combinazione delle due. GSD e ESD non sono mutualmente esclusive e sono state identificate molte specie di vertebrati in cui entrambi i meccanismi operano simultaneamente in risposta ad un continuum di fattori ereditari e ambientali (Capel, 2017).

Nella GSD, il sesso primario di un individuo è determinato al momento della formazione dello zigote (Li e Gui, 2018). In mammiferi, uccelli, rettili, anfibi e pesci è stata riconosciuta una classificazione attraverso la differenziazione dei cromosomi sessuali in un sistema eterogametico maschile XY ed in un sistema eterogametico femminile ZW, (Pennell, et al., 2018).

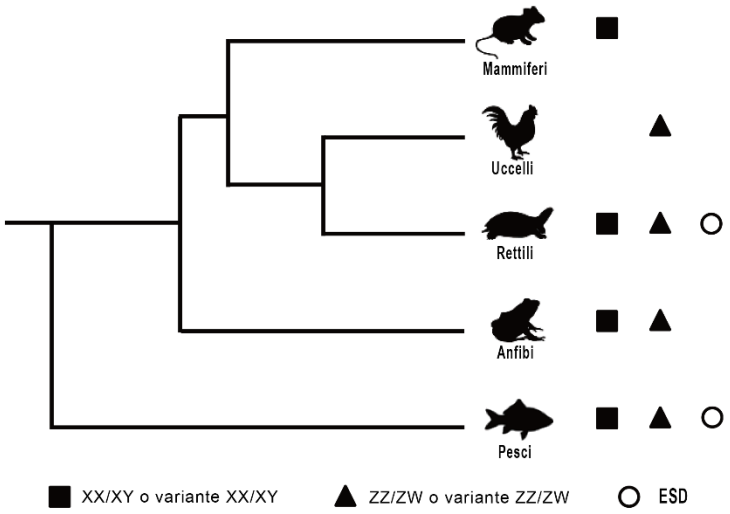

Figura 1. Schematizzazione dei diversi sistemi di determinazione del sesso identificati nei vertebrati. "Variante" indica l'assenza di un cromosoma sessuale o la presenza di più cromosomi sessuali. ESD, determinazione del sesso ambientale (environmental sex determination), tratto da Li e Gui (2018), modificato.

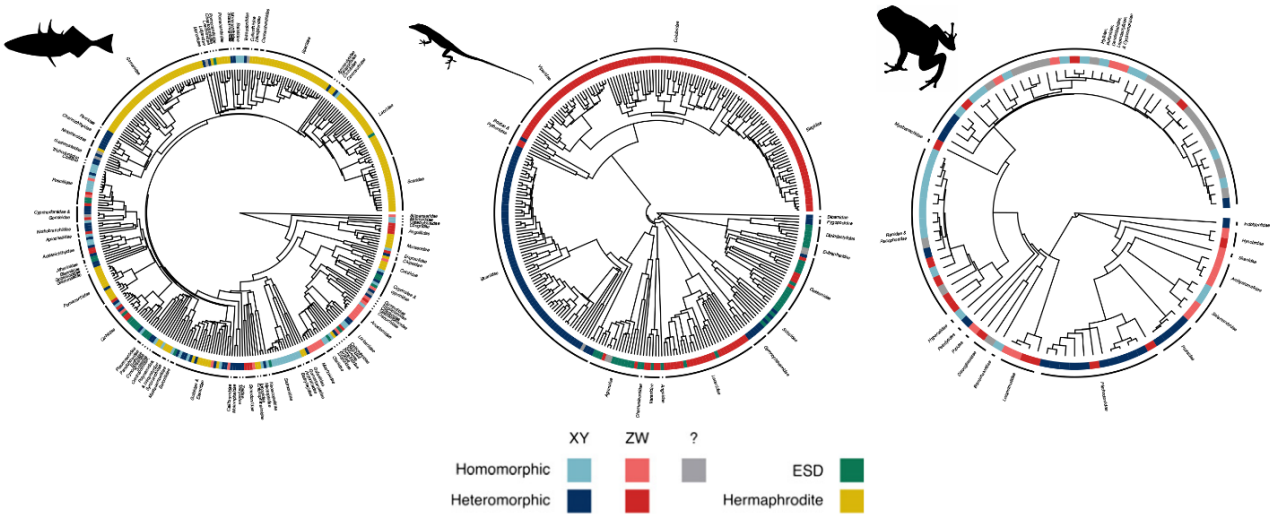

Figura 2. Distribuzione della determinazione del sesso in tre cladi di vertebrati. Le specie sono codificate come XY eteromorfe (blu scuro), XY omomorfe (azzurro), ZW eteromorfe (rosso scuro), ZW omomorfe (rosso chiaro), omomorfe sconosciuto (?, grigio), determinazione del sesso ambientale (ESD, verde) o ermafroditi (giallo). (Pennell, et al., 2018)

Sebbene la GSD possa essere poligenica (Moore e Roberts, 2013) è più frequentemente sotto il controllo di un unico locus, risultando in un sistema dove il maschio o la femmina sono eterozigoti al locus determinante il sesso (Pennell, et al., 2018).

I meccanismi che controllano il modo in cui viene determinato il sesso differiscono considerevolmente tra i vari cladi, si sono evoluti ripetutamente e indipendentemente e i percorsi molecolari sottostanti possono cambiare rapidamente durante l'evoluzione (Herpin e Schartl, 2015). Sono stati tuttavia identificati negli anni degli attori principali, denominati "*master sex-determining genes*" (MSD *genes*), che derivano comunemente da 3 tipologie di geni: i) fattori di trascrizione appartenenti alla famiglia proteica "DM *domain*" (*Drosophila melanogaster doublesex proteins*), ii) geni appartenenti alla cascata di segnale del TGF-β e iii) geni della famiglia *sox* (*Sry-related HMG box*). Principalmente nei teleostei, hanno poi acquisito il ruolo di geni MSD anche geni già appartenenti alla via di regolazione sessuale (i.e. sox3y, hsd17b1) o geni immuno-correlati (i.e. irf9).

Gene determinante il sesso	**Clade/ specie**	**Sistema di determinazione sessuale**	**Classe**	**Riferimenti**
Sry	Mammiferi teri	XY	Mammiferi	(Koopman, et al., 1991)
sox3y	*Oryzias dancena*	XY	Pesci	(Takehana, et al., 2014)
Dmrt1	Uccelli	ZW	Uccelli	(Smith, et al., 2009)
dmrt1	*Cynoglossus semilaevis*	ZW	Pesci	(Chen, et al., 2014)
Dmrt1	*Trachemys scripta*	TSD	Rettili	(Ge, et al., 2017)
dmy	*Oryzias latipes*	XY	Pesci	(Matsuda, 2005)
Dm-w	*Xenopus laevis*	ZW	Anfibi	(Yoshimoto, et al., 2008)
Amhy	*Ornithorhynchus anatinus*	$X_1X_2X_3X_4X_5$ $Y_1Y_2Y_3Y_4Y_5$	Mammiferi	(Cortez, et al., 2014)
amhy	*Odontesthes hatcheri*	XY	Pesci	(Hattori, et al., 2012)
amhy	*Oreochromis niloticus*	XY	Pesci	(Li, et al., 2015)
amhr2	*Takifugu rubripe*	XY	Pesci	(Kamiya, et al., 2012)
gsdfy	*Oryzias luzonensis*	XY	Pesci	(Myosho, et al., 2012)
gsdfy	*Nothobranchius furzeri*	XY	Pesci	(Reichwald, et al., 2015)
sdy	*Oncorhynchus mykiss*	XY	Pesci	(Yano, et al., 2012)

Tabella 1. Geni determinanti il sesso o candidati nei vertebrati, tratta da Li e Gui (2018), modificata.

1.1.1. Mammiferi

Nei mammiferi il sistema prevalente di differenziazione sessuale è quello eterogametico maschile XY, in cui gli individui di sesso femminile possiedono due cromosomi X, mentre quelli di sesso maschile possiedono un cromosoma X ed uno Y. La trasmissione paterna

alla prole di un cromosoma Y innesca la differenziazione testicolare, mentre la presenza nel gamete paterno del cromosoma X spinge le gonadi verso la differenziazione ovarica.

Il primo gene identificato come determinante del sesso in molti mammiferi (tra cui anche nell'uomo) è stato il gene *Sry* (Sex-determining Region Y) (Sinclair, et al., 1990) , localizzato sul cromosoma Y. La determinazione sessuale maschile inizia grazie all'espressione di *Sry*, che determina l'espressione della proteina *Sox9* e, a sua volta, di *Fgf9*. Questo aumenta la sintesi dell'ormone *prostaglandina* D_2 che aiuta a mantenere l'espressione di *Sox9* portando alla creazione di un feedback positivo e aiutando la differenziazione cellulare gonadica delle cellule del Sertoli. Allo stesso tempo, *Sox9* e *Fgf9* sopprimono la cascata di segnali che determinano la differenziazione ovarica (Nef e Vassalli, 2009).

Nonostante la maggior parte dei mammiferi dipendano dal sistema XY con *Sry* all'apice della cascata determinante il sesso, sono state riscontrate varianti del sistema XX/XY in alcune specie, come in *Tokudaia osimensis osimensis* e *Tokudaia osimensis spp. ,* piccoli roditori in cui è stato perso il cromosoma Y (e di conseguenza il gene *Sry*). Sia maschi che femmine hanno di conseguenza un identico cariotipo XO (Sutou, et al., 2001) e non è ancora stato riconosciuto quali siano i geni (o il gene) che determinino il sesso in questi mammiferi in assenza di *Sry.*

Un'ulteriore eccezione è stata riscontrata in alcune specie di monotremi. In particolare, negli ornitorinchi (*Ornithorhynchus anatinus*) sono presenti 5 paia di cromosomi X negli individui di sesso femminile e 5 cromosomi X e 5 cromosomi Y in quelli di sesso maschile (Rens, et al., 2004). Cariotipo simile è stato identificato nell'echidna istrice (*Tachyglossus aculeatus*) che nel sesso maschile presenta 5 cromosomi X e 4 cromosomi Y (Rens, et al., 2007). Questi due mammiferi non hanno il gene *Sry* (Wallis, et al., 2007) e il più probabile gene determinante il sesso è il gene *Amh* presente nel cromosoma Y_5 (Cortez, et al., 2014) . Il gene *Amh* codifica l'ormone *anti-Mülleriano* (AMH) prodotto dalle cellule del Sertoli, il quale, legandosi ai recettori presenti nei dotti di Müller, induce l'apoptosi degli stessi causandone la regressione e impedendo lo sviluppo dell'utero e delle salpingi. (Chadwick e Goode, 2002).

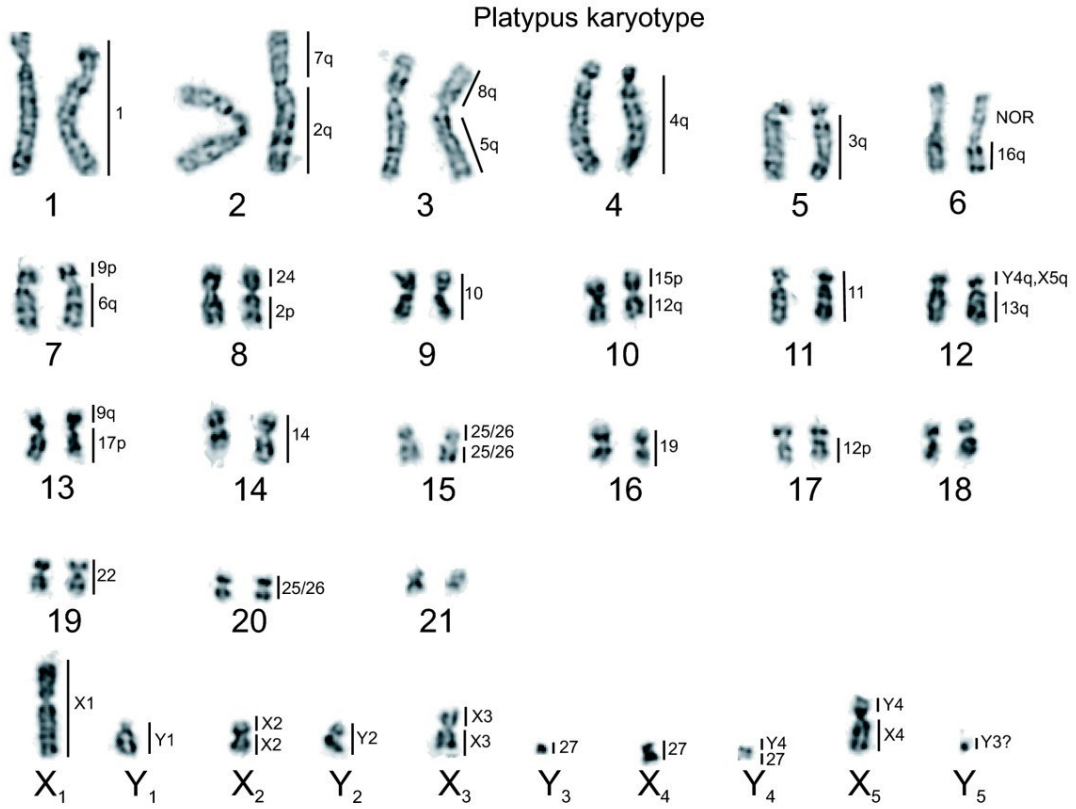

Figura 3. Cariotipo di Ornithorhynchus anatinus, tratto da Rens et al. (2007). I numeri sulla destra rappresentano le regioni omologhe tra Ornithorhynchus anatinus e Tachyglossus aculeatus

1.1.2. Uccelli

A differenza dei mammiferi, gli uccelli presentano un sistema di determinazione del sesso di tipo ZZ/ZW, con cromosomi eterogametici per il sesso femminile: l'individuo femminile presenta un cromosoma Z e uno W mentre un individuo maschile presenta due cromosomi Z.

Anche nel caso degli uccelli, diversi geni sono stati proposti come candidati per la differenziazione del sesso. Tra questi, il primo ad essere stato confermato è stato *Dmrt1* (localizzato sul cromosoma Z) nel pollo (Lambeth, et al., 2014).

La sovra espressione di *Dmrt1* induce l'attivazione dell'espressione di SOX9 e la soppressione dell'enzima aromatasi (deputato alla conversione degli androgeni in estrogeni) (Smith, et al., 2009). Questo, di conseguenza, induce una differenziazione sessuale maschile e antagonizza il differenziamento femminile delle gonadi. Nel caso in cui la presenza di *Dmrt1* fosse esigua (come nel caso di un cariotipo ZW), l'espressione di *Foxl2* viene inibita inducendo l'attivazione dell'enzima aromatasi, così da portare allo sviluppo femminile (Sánchez e Chaouiya, 2018).

Il gene *dmrt1* è stato confermato essere coinvolto nella differenziazione sessuale non solo negli uccelli, ma anche in alcuni pesci (Matsuda, et al., 2002), rettili (Murdock e Wibbels, 2006) e mammiferi (Smith, et al., 1999).

1.1.3. Anfibi

Negli anfibi sono stati riscontrate sia specie con determinazione del sesso di tipo XX/XY che di tipo ZZ/ZW. Nelle specie con eterozigosi maschile le degenerazioni del cromosoma Y sono talmente limitate da rendere, nella maggior parte dei casi, morfologicamente indistinguibili i due cromosomi (Gamble, et al., 2015) : il numero di specie in cui è stata rilevata una differenza identificabile con tecniche citogenetiche è stato solamente di 20 su 1500 specie studiate (Eggert, 2004).

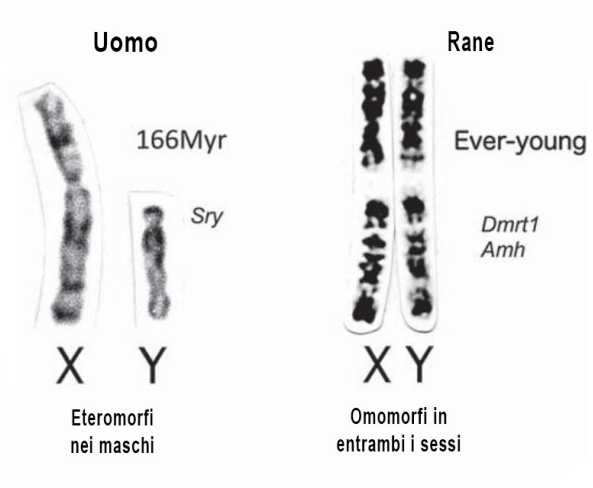

Figura 4. Cromosomi sessuali umani e di rana. I cromosomi X e Y sono eteromorfi nell'uomo, mentre sono omomorfi in entrambi i sessi delle rane. Le fotografie dei cromosomi sessuali sono state fornite da Chizuko Nishida. (Miura, 2017)

Nelle specie studiate ad oggi, il sesso è determinato geneticamente (GSD), non sono stati identificati né una determinazione legata alla temperatura ("determinante" fra i più frequente tra gli ESD) né altre determinazioni sessuali ambientali; tuttavia, è possibile far variare il sesso fenotipico in alcune specie attraverso la temperatura o la somministrazione esogena di ormoni steroidei e i loro inibitori (Miura, 2017).

Nelle rane sono stati proposti 8 differenti geni, distribuiti su 6 differenti cromosomi, per essere i geni deputati alla determinazione sessuale: *Dmrt1, Amh, Cyp19, FoxL2, Sox3, Sf1, Ar, Cyp17* (Miura, 2017).

Nella rana *Xenopus laevis,* in cui sono presenti cromosomi omomorfi di tipo ZZ/ZW, è stato identificato nel cromosoma W un gene con un'identità pari all'89% rispetto al gene *Dmrt1,* chiamato *dm-w*, il cui ruolo è fondamentale per lo sviluppo dell'ovaio ed è quindi un candidato per la determinazione del sesso femminile (Yoshimoto, et al., 2008).

1.1.4. Rettili

L'elenco di geni MSD (*master sex-determining*) validati non include finora specie di rettili. L'insieme di molti fattori ha portato alla sottovalutazione della presenza di geni determinanti il sesso nei rettili, anche a causa di una minore attenzione da parte dei ricercatori per la determinazione genetica del sesso rispetto alla determinazione del

sesso ambientale presente in questa classe di vertebrati. La determinazione del sesso attraverso la temperatura nei rettili è una caratteristica peculiare e si sono concentrati maggiori sforzi in questo ambito a scapito della determinazione genetica. Inoltre, l'assenza di cromosomi eteromorfi ha reso più complesso l'identificazione e lo studio dei cromosomi sessuali in questa classe. Infine, un'ulteriore causa è l'assenza di interesse economico, in quanto nessun rettile ha una rilevanza commerciale così elevata da spingere la ricerca nello sviluppo di nuove tecnologie. (Thépot, 2021)

Tutti questi motivi spiegano perché l'identificazione dei geni MSD nei rettili sia ancora in ritardo rispetto alla ricerca in altri gruppi. Ad ogni modo, con lo sviluppo di nuove tecnologie, questa lacuna sta cominciando a essere colmata e sono stati compiuti incredibili progressi durante gli ultimi due decenni sebbene il livello di conoscenza sia ancora molto eterogeneo nelle differenti famiglie (Thépot, 2021).

Nei rettili sono stati riconosciute sia specie XX/XY che ZZ/ZW, tali da suggerire l'esistenza di geni determinanti il sesso: motivo per cui sono stati ricercati geni già riconosciuti essere determinanti il sesso in altre specie. Ad esempio, nella tartaruga dalle orecchie rosse (*Trachemys scripta), Dmrt1* esibisce un'espressione temperatura-dipendente differente nei due sessi, ed è un importante candidato per essere un gene determinante il sesso anche in questa specie (Ge, et al., 2017).

1.1.5. Pesci

Più della metà delle specie di vertebrati sono pesci, con oltre 32'000 specie riconosciute in tutto il mondo (Nelson, et al., 2016). I pesci teleostei dimostrano una grande varietà di meccanismi di determinazione del sesso che coinvolgono fattori genetici, ambientali e sociali. Le modalità di differenziazione sessuale variano dall'ermafroditismo al gonocorismo e, nonostante siano state identificate specie gonocoriche, sono poche le specie in cui sono stati identificati cromosomi eteromorfi (Arkhipchuk, 1995).

Nella maggior parte delle specie gonocoriche, la gonade indifferenziata bipotente sviluppa direttamente in un ovario o testicolo (con l'eccezione dell'ermafroditismo giovanile), mentre le specie ermafrodite possono essere sincroniche (in cui tessuto ovarico e testicolare sono simultaneamente attivi nello stesso organismo) o sequenziali (tessuto ovarico e testicolare si alternano nell'organismo, con inversione da maschio a femmina nelle specie proterandriche o con inversione da femmina a maschio nelle specie proteroginiche) (Sadovy e Shapiro, 1987).

Allo stesso tempo, nonostante in alcune specie ermafrodite sia presente un rilevante fattore genetico, alcuni fattori ambientali (come la temperatura, il pH e determinate

condizioni sociali) possono influenzare la determinazione del sesso e sono riconosciute sempre più specie in cui possono susseguirsi molteplici inversioni sessuali. Un esempio è *Trimma okinawae,* in cui il cambiamento da femmina a maschio avviene quando l'individuo diventa il più grande del suo gruppo. I maschi possono poi ritornare ad essere femmine se cambiano gruppo e trovano un individuo maschio di dimensioni maggiori rispetto alla loro (Munday, et al., 2006).

Per quanto riguarda le specie gonocoriche con cromosomi eteromorfi, oltre che i cromosomi XX/XY e ZZ/ZW, sono stati identificati XX/XO, ZZ/ZO, XX/XYY_2, $X_1X_2X_1X_2/X_1X_2Y$ (Chen, et al., 2022) ma non è stato ancora possibile identificare un gene universale equivalente allo *Sry* dei mammiferi.

Ad esempio, in medaka (*Oryzias latipes*), un piccolo pesce eterogametico ma con cromosomi X e Y non citologicamente differenziabili, è stato identificato un gene determinante il sesso nel cromosoma Y: *dm* correlato a PG 17, denominato *dmy* perché specifico del cromosoma Y. È stata inoltre suggerita la necessità dell'espressione di dmy per un normale sviluppo testicolare, in quanto in individui con mutazioni a livello di *dmy* e con genotipo XY risultavano fenotipicamente femmine (Matsuda, et al., 2002).

Tuttavia, questo gene è stato riportato solamente in *Oryzias latipes* e *Oryzias curvinotus*, (Matsuda, et al., 2003) e non è stato ritrovato in altre specie del genere *Oryzias* (Kondo, et al., 2003). In *Oryzias dancena* è stato invece identificato il gene *sox3y,* responsabile dell'inizio della differenziazione testicolare sovra regolando l'espressione di *Gsdf*, un gene chiave per la differenziazione testicolare nei teleostei (Takehana, et al., 2014).

Dalla scoperta del *dmy*, nessun nuovo gene determinante il sesso è stato identificato nei pesci per dieci anni, a causa della mancanza di metodi efficaci e di materiali idonei. Nell'ultimo decennio, lo sviluppo di tecnologie di sequenziamento NGS (Next Generation Sequencing), rapide e ad elevato *througput*, è diventato il motore per la scoperta di nuovi geni determinanti il sesso nei pesci. Inoltre, la tecnologia di *editing* genomico ha fornito un notevole supporto tecnico per l'identificazione di geni candidati. Tra il 2012 e il 2015, in soli tre anni, sono stati scoperti oltre 7 geni MSD (Chen, et al., 2022).

Ad oggi, ci sono 2 principali strategie per isolare i geni determinanti il sesso con le tecnologie NGS: confrontare i trascritti sessuali specifici nella fase della differenziazione gonadica con il sequenziamento del trascrittoma oppure eseguire lo screening di loci di DNA specifici del sesso o legati al sesso confrontando i genomi maschili e femminili per identificare i geni/allei candidati che determinano il sesso in base alla loro presenza nel genoma specifico del sesso d'interesse (Chen, et al., 2022).

Utilizzando queste tecniche, nell'ultimo decennio sono numerosi i geni che sono stati riscontrati come possibili determinanti il sesso nei pesci, alcuni già riscontrati in classi differenti, tra cui *dmrt1* presente nel cromosoma Z in *Cynoglossus semilaevis* (Chen, et al., 2014), altri completamente differenti, come *amh* in *Odontesthes hatcheri*.

Odontesthes hatcheri è un pesce gonocorico con un sistema di determinazione sessuale di tipo XX/XY, in cui è stata riscontrata la presenza del gene *amh (*strettamente legato al cromosoma y, per cui chiamato *amhy* - Y chromosome-specific *amh)* pur non avendo i dotti di Müller, target dell'ormone AMH nei mammiferi. È stato inoltre dimostrato che *amhy* è richiesto per lo sviluppo testicolare suggerendo così che un gene correlato ad un ormone è un possibile meccanismo alternativo per il controllo trascrizionale della determinazione del sesso nei teleostei (Hattori, et al., 2012).

1.2. *Mugil cephalus*

1.2.1. Areale e inquadramento sistematico

Il cefalo (*Mugil cephalus*) è un pesce osseo appartenente alla famiglia mugilidae. Il suo habitat consiste nelle acque costiere e negli estuari delle zone tropicali e subtropicali di tutti i mari. La distribuzione è all'incirca tra 42° N. e 42° S. Ad eccezione del Mar Mediterraneo e del Mar Nero, questa specie non sembra abitare acque in cui la temperatura media mensile dell'acqua scenda al di sotto di 16° C o dove la temperatura estiva non sia superiore ai 18° C (Thomson, 1963).

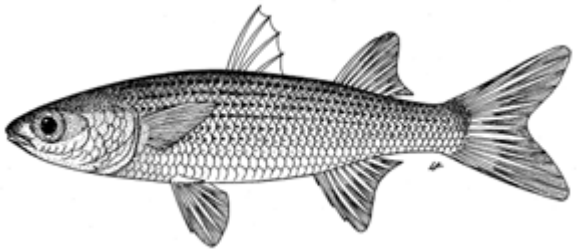

Figura 5. Mugil cephalus (www.fao.org)

1.2.2. Caratteristiche riproduttive

Il cefalo è catadromo, trascorre la maggior parte della sua vita nelle acque costiere degli estuari e negli ambienti di acqua dolce, per poi migrare al largo per deporre le uova. La riproduzione avviene in mare, in acque superficiali e solitamente durante le ore notturne. Il periodo riproduttivo varia a seconda dell'areale in entrambi gli emisferi, ad esempio in Italia i riproduttori migrano dalle acque interne al mare in agosto, e la riproduzione avviene entro settembre.

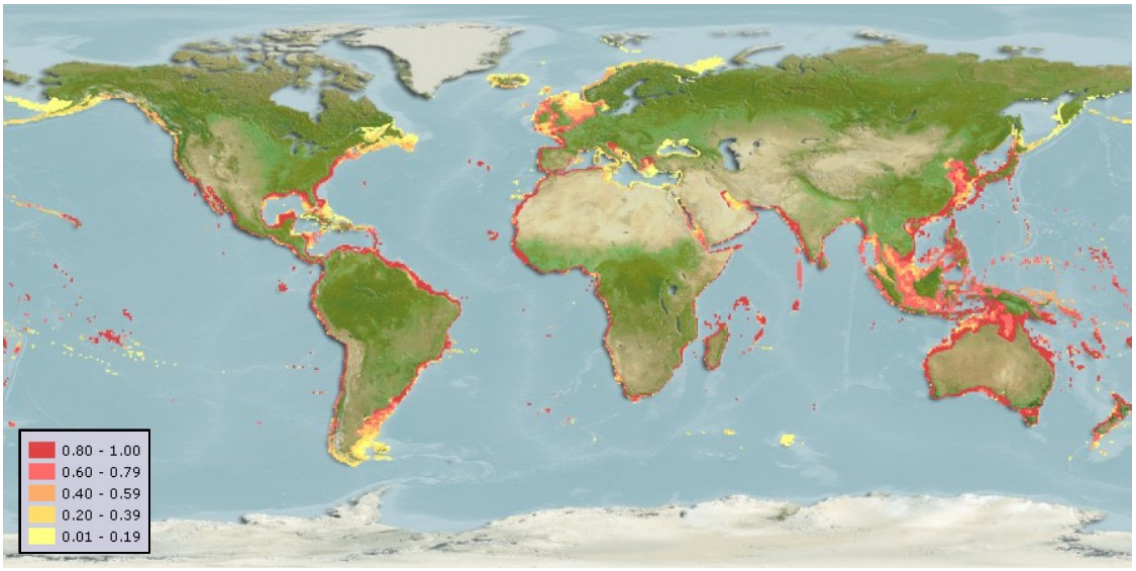

Figura 6. Areale di Mugil cephalus. I colori dell'intervallo di distribuzione indicano la probabilità con cui un determinato ambiente può essere considerato habitat per Mugil cephalus. (www.fishbase.se)

La fecondità è alta e il numero di uova deposte da una femmina è direttamente correlato alla sua taglia. Sono presenti anche alcune femmine che depongono solo ad anni alterni dopo la prima maturità. Ogni individuo può deporre oltre 4.800.000 uova di colore giallo paglierino e dal diametro di circa 1,08 mm, con all'interno una grande goccia oleosa che ne favorisce il galleggiamento (http://www.ittiofauna.org/).

I cefali adulti sono stati trovati in acque che vanno da salinità zero a 75‰, mentre i giovani possono tollerare intervalli di salinità così ampi solo dopo aver raggiunto lunghezze di 4-7 cm. Le larve si spostano verso la costa in acque estremamente basse, che forniscono copertura dai predatori e un ricco terreno di alimentazione. Dopo aver raggiunto i 5 cm di lunghezza, i giovani cefali si spostano in acque leggermente più profonde (https://www.fao.org/).

Il dimorfismo sessuale nei cefali non è evidente: le femmine sono generalmente più grandi dei maschi ma è molto complessa la distinzione nei due sessi prima della completa maturità sessuale (http://www.ittiofauna.org/).

La maturità sessuale viene in genere raggiunta tra i 2 e i 4 anni, (Whitfield, et al., 2012) con variazioni legate all'origine geografica. Presso la laguna di Tortolì è stata stimata un'età alla cattura tra i 4 e 9 anni (Diciotti, et al., 2022).

1.2.3. Interesse commerciale

Il cefalo è molto popolare sia nella pesca che nell'acquacultura, soprattutto in paesi come il Giappone, Taiwan e Korea e in alcune aree del Mar Mediterraneo (Mar Tirreno, nella costa Ovest della penisola italiana e nelle pozioni più a nord ed Ovest della Grecia).

In particolare, nelle aree del mediterraneo ha un importante valore commerciale per l'economia locale grazie alla produzione della bottarga (Bekhit, 2022).

La bottarga è una specialità culinaria costituita dall'ovario del muggine, pressato ed essiccato, ed è considerato un cibo molto pregiato. In molte parti del mondo ha destato notevole interesse negli ultimi decenni, tanto che il cefalo viene anche chiamato "oro grigio" dai pescatori (Shaw e Daigee, 2006).

1.2.4. Marcatori sessuali identificati

M. cephalus è una specie gonocorica con determinazione genetica del sesso sebbene caratteristiche ermafrodite non funzionali siano state visualizzate in gonadi mature differenziate (McDonough, et al., 2005). Il cariotipo del cefalo contiene 24 coppie di cromosomi ed è stata provata l'assenza di cromosomi sessuali eterocromatici (Rossi, et al., 1996).

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; pertanto, in questa specie è possibile identificare marcatori genetici associati al sesso.

In un recente studio, condotto dal gruppo di ricerca presso cui ho svolto il mio lavoro di tesi, (Ferraresso, et al., 2021) è stato utilizzato un approccio di sequenziamento dell'intero genoma, tramite Pool-Seq, con l'obiettivo di identificare marcatori genetici legati al sesso.

Un pool di 60 femmine e un pool di 60 maschi, (ciascuno composto da individui appartenenti a due popolazioni sarde, Cabras-CAB e Tortolì-TOR) sono stati sequenziati con tecnologia NGS per l'identificazione di varianti nucleotidiche (SNP) associabili al sesso. Tale analisi ha permesso l'identificazione di 3 SNP localizzati nel gene *fshr* che mostrano una frequenza tale da poter essere considerati sesso-specifici (loci che manifestano un singolo allele in un sesso e 2 alleli nell'altro sesso). Le tre varianti (denominate MuCe179, MuCe206 e MuCe322 in relazione alla loro posizione nel genoma di riferimento) sono risultate essere localizzate nell' esone 14 del gene *fshr*. MuCe179 e MuCe206 sono mutazioni missenso, mentre MuCe322 è una mutazione sinonima.

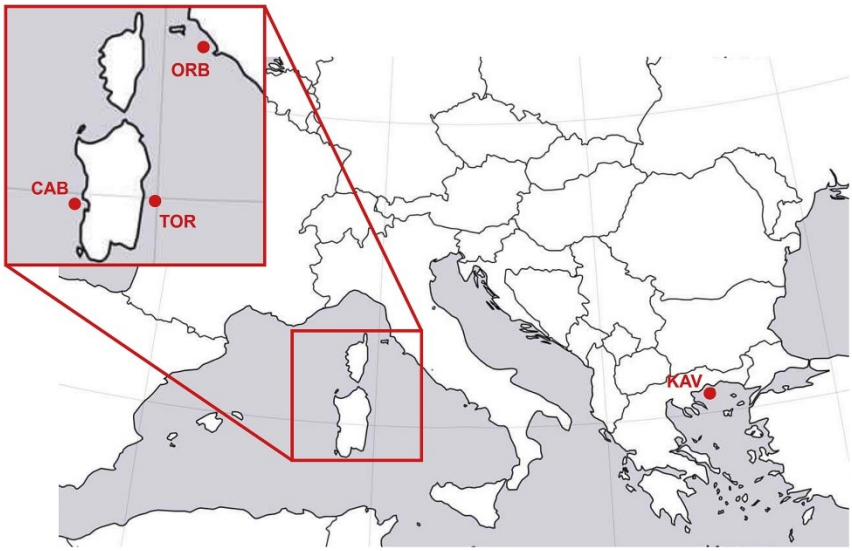

Figura 8. Localizzazione geografica delle 4 popolazioni mediterranee di Mugil cephalus studiate (Ferraresso, et al., 2021).

Il gene *fshr* codifica per i recettori dell'ormone follicolo-stimolante (FSHR). L'ormone follicolo-stimolante (FSH) viene secreto dall'ipofisi e, nei vertebrati, si lega ai recettori localizzati sulle cellule della granulosa nell'ovaio e sulle cellule del Sertoli nel testicolo per regolare lo sviluppo delle gonadi e promuoverne la crescita. Anche nei pesci è stata confermata la presenza di FSHR nelle cellule della granulosa e nelle cellule del Sertoli e si ipotizza che l'espressione di *fshr* nelle cellule del Sertoli di zebrafish possa essere legato al gonocorismo indifferenziato (García-López, et al., 2010). In uno studio (Zhang, et al., 2015) a seguito dell'ottenimento di zebrafish knock-out (KO) per *fshr* le femmine hanno mostrato un completo fallimento dell'attivazione follicolare e hanno subito un *sex reversal* verso il sesso maschile, mostrando una normale fertilità.

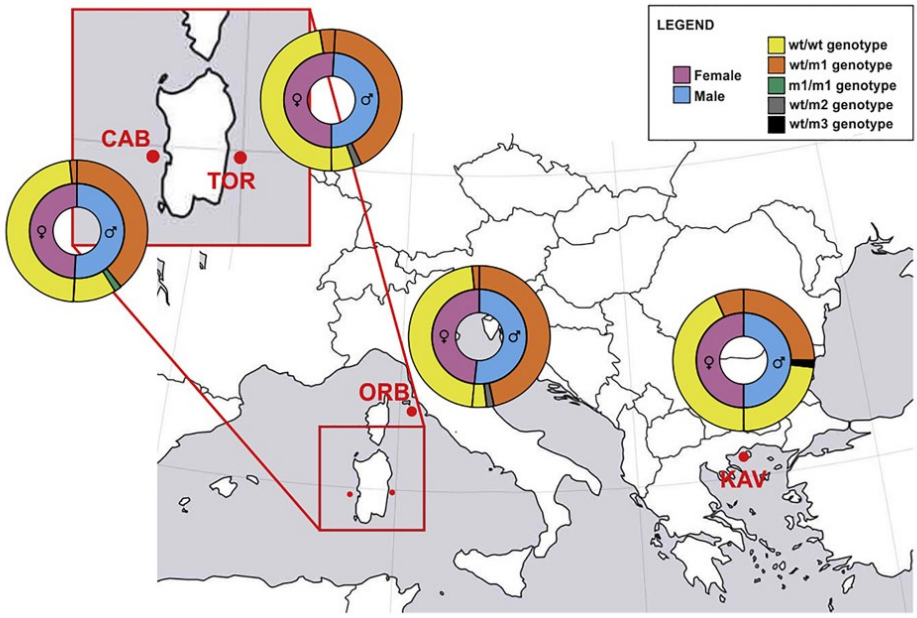

Figura 7. Correlazione tra genotipi e sesso nelle differenti popolazioni del mediterraneo (Ferraresso, et al., 2021)

Nei maschi, in assenza di *fshr*, la spermatogenesi era normale negli adulti anche se è stato riscontrato un ritardo nella maturità sessuale dei giovani.

Per validare i risultati ottenuti con il Pool-Seq, una regione di 308 nucleotidi che comprende MuCe179, MuCe206 e MuCe322 è stata sequenziata in 245 individui, provenienti da 3 popolazioni del Tirreno (Laguna di Cabras – CAB; Laguna di Orbetello – ORB; Tortolì – TOR), e una popolazione del Mar Egeo settentrionale (Baia di Kavala – KAV).

L'allele più frequente è stato considerato wild type (wt), m1 è stato chiamato l'allele che comprende le tre varianti (MuCe179, MuCe206 e MuCe322) mentre m2 è stato chiamato un secondo allele contenente le 2 mutazioni *missenso* (MuCe179 e MuCe206).

I dati relativi al genotipo dell'*fshr* hanno confermato che il genotipo wt/wt è significativamente associato al sesso femminile. È risultato inoltre evidente che le due mutazioni *missenso* sono significativamente associate al sesso maschile e il pattern wt/m1 e wt/m2 dei maschi suggerisce un tipo di gene determinante il sesso di tipo "XX-XY". I due SNP, presenti in eterozigosi (wt/m1 o wt/m2) nel fenotipo maschile, hanno però mostrato penetranza incompleta in un certo numero di maschi, definiti "non-conformi", che esibiscono un genotipo wt/wt, associato invece al fenotipo femminile. La frequenza di maschi non-conformi si è mostrata relativamente bassa (6-18%) nelle popolazioni italiane analizzate, mentre ha raggiunto una frequenza vicina al 50% nella popolazione dell'Egeo. Questo suggerisce il possibile coinvolgimento anche di altri loci nella determinazione del sesso nel cefalo e/o il possibile ruolo anche di fattori ambientali.

Recentemente è stato identificato anche in *Solea senegalensis* una correlazione tra *fshr* e il sesso, identificando 41 varianti alleliche nel gene dell'*fshr* e riscontrando omozigosi nelle femmine ed eterozigosi nei maschi, concordi con il sistema di determinazione sessuale di tipo XX/XY (Herrán, et al., 2022).

1.3. Sequenziamento di nuova generazione (NGS)

1.3.1. Tecniche di sequenziamento

Con sequenziamento si intende il riconoscimento e la determinazione dell'esatto ordine dei nucleotidi che compongono il DNA di un individuo.

Nel corso degli anni si sono susseguite diverse strategie per determinare la sequenza nucleotidica di un frammento di DNA. La prima metodica, il metodo di sequenziamento Sanger, risale al 1977 ed è alla base delle metodiche di prima generazione, in cui i frammenti vengono separati in base alla loro dimensione e analizzati con l'elettroforesi su gel per determinarne l'esatta sequenza (Sanger, et al., 1977). Il metodo Sanger ha poi continuato a migliorarsi con l'introduzione dell'elettroforesi capillare ed è stato utilizzato per sequenziare differenti tipologie di genomi, da quelli di batteri e fagi fino ad arrivare al genoma umano, lavoro per cui è stato necessaria più di una decade e 2,7 miliardi di dollari (Hu, et al., 2021) rendendo evidente la necessità di nuove tecnologie per poter continuare a sequenziare in modo rapido ed efficiente.

All'inizio degli anni 2000 sono state introdotte le tecnologie di seconda generazione (*next-generation sequencing*, NGS) basate sul sequenziamento in parallelo di milioni di frammenti amplificati di DNA di piccole dimensioni (60-300 nucleotidi). Generalmente ci si riferisce alle piattaforme Illumina o Ion Torrent, il cui flusso di lavoro comprende la preparazione di librerie, il sequenziamento e l'analisi dei dati (Hu, et al., 2021).

Grazie a queste tecnologie, negli ultimi decenni c'è stata una massiva riduzione dei costi per il sequenziamento di interi genomi ed i punti fondamentali che molti studiosi oggi si trovano ad affrontare sono sulle decisioni riguardanti quanto genoma sequenziare (*breadth of coverage*), quanto andare nel dettaglio per ogni campione (*depth of coverage*) e quanti campioni sequenziare (Lou, et al., 2021).

Il sequenziamento di terza generazione invece può eseguire letture di lunghezza fino a 10kb e per questo motivo vengono chiamate tecnologie "*long read*" (alcuni esempi sono la piattaforma Pacific Biosciences e Oxford Nanopore). Sebbene riescano a superare alcuni problemi riscontrati nella seconda generazione, quali sequenze ripetute all'interno di un intero genoma, e richiedano passaggi minimi di preparazione della libreria, 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 (Hu, et al., 2021).

1.3.2. Tecnologia Illumina

La tecnologia di sequenziamento Illumina è nota come sequenziamento mediante sintesi (SBS) e prevede tre passaggi principali: preparazione delle librerie, generazione di clusters e sequenziamento.

Nella preparazione delle librerie il DNA viene frammentato in segmenti di una data lunghezza a cui verranno aggiunti adattatori specifici ad entrambe le estremità.

Il costrutto finale contiene gli adattatori (sequenze che consentono alla libreria di legarsi e generare cluster sulla *flowcell* - sequenze p5 e p7), le sequenze dei siti di legame del primer per avviare il sequenziamento (Rd1 SP e Rd2 SP) e le sequenze degli identificatori univoci di ciascun campione (*index* 1 e, ove applicabile, *index* 2) che consentono il *multiplexing/pooling* di più campioni in una singola corsa di sequenziamento.

Figura 9. Rappresentazione schematica di un frammento di libreria con due indici

Sarà in seguito necessaria un'amplificazione delle sequenze, in modo da fornire un segnale sufficiente per il sequenziamento.

La generazione di cluster è il processo in base al quale ciascun frammento in una libreria viene clonato in migliaia di copie identiche. Per generare i *cluster*, la libreria viene caricata in una *flowcell* (un vetrino dotato di diverse corsie la cui superficie è ricoperta da due tipi di oligonucleotidi, complementari ai due differenti adattatori presenti sulle librerie), in cui i frammenti si ibridano e una polimerasi creando un complemento del frammento ibridato. La molecola a doppio filamento viene denaturata e il campione d'origine viene dilavato. Ogni frammento, legato per un'estremità alla *flowcell*, si piega e si lega ai primer adiacenti, formando un ponte, da qui il nome *bridge amplification*. Una polimerasi sintetizza il frammento *reverse* e tale processo di amplificazione viene ripetuto più volte, così che a seguito di cicli consecutivi di denaturazione, appaiamento ed estensione si abbia la formazione di *cluster* contenenti fino ad un migliaio di copie identiche alla molecola iniziale.

Figura 10. Rappresentazione schematica della generazione di clusters (www.illumina.com).

Dopo l'amplificazione a ponte i frammenti *reverse* vengono tagliati e dilavati, lasciando solo i frammenti *forward*, che vengono linearizzati per preparali al sequenziamento.

Per il sequenziamento viene sfruttata la tecnologia di sequenziamento mediante sintesi (SBS). Il sequenziamento inizia con l'estensione del primo primer di sequenziamento per produrre la prima *read.* Ad ogni ciclo, i nucleotidi marcati mediante fluorofori competono per essere addizionati alla catena in crescita, e ne viene incorporato solamente uno, complementare alla sequenza modello. Dopo l'aggiunta di ogni nucleotide, i cluster vengono eccitati da una sorgente luminosa e ciascuna delle basi emette una lunghezza d'onda univoca che la renderà identificabile.

Dopo il completamento della prima *read*, il prodotto letto viene dilavato. Viene successivamente introdotto il primer del primo indice e ibridato al modello, così da generare una lettura simile alla prima *read.* Dopo il completamento della lettura dell'indice il prodotto letto viene dilavato e l'estremità 3' del frammento modello viene deprotetta. Il modello ora si piega nuovamente e il secondo indice viene letto alla stessa maniera del primo. Le polimerasi estendono il secondo oligonucleotide della cella a flusso formando un ponte a doppio filamento, il quale viene linearizzato e vengono bloccate le estremità 3'. Il frammento originale *forward* viene dilavato lasciando solo il frammento *reverse*. La *read* 2 inizia con l'introduzione di un primer apposito e, come per la *read* 1, i passaggi vengono ripetuti fino al raggiungimento della lunghezza desiderata. L'approccio *paired-end* (PE) *sequencing*, permette di sequenziare a partire da entrambe le estremità dei frammenti di DNA e di ottenere *reads forward* e *reverse* (R1 e R2), per un risultato più accurato.

Nella fase di sequenziamento vera e propria, esistono 3 diverse metodologie per differenziare le differenti basi, con un tipo di chimica a quattro canali, due canali o un canale, a seconda della piattaforma utilizzata. Nel presente lavoro di tesi, lo strumento utilizzato per il sequenziamento Novaseq6000 utilizza la chimica a due canali.

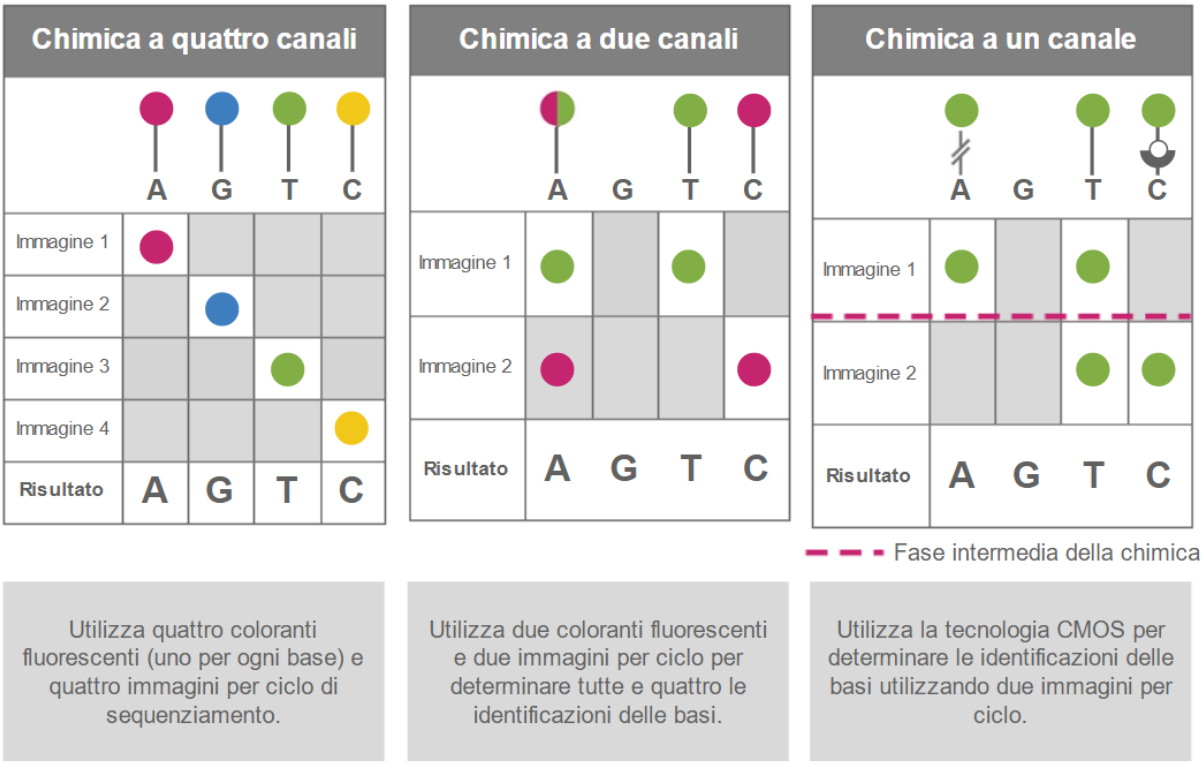

Figura 11. Confronto delle differenti tipologie di chimica a canali Illumina (www.Illumina.com)

1.3.3. Variant calling from whole genome sequencing

Le varianti a singolo nucleotide, chiamate più comunemente SNP (*Single Nucleotide Polymorphism*) sono un tipo di polimorfismo genetico nel quale un determinato gene presenta, in individui diversi, variazioni di sequenza a carico di una singola base della catena polinucleotidica.

Mutazioni puntiformi possono presentarsi all' interno di un gene, sia nelle regioni codificanti che in quelle non codificanti, o in una regione intergenica.

Gli SNP presenti in regioni codificanti possono essere classificati come sostituzioni sinonime (non c'è una variazione la sequenza amminoacidica codificata in quanto - a seguito della "ridondanza del codice genetico" - lo stesso amminoacido può essere codificato da più sequenze di basi/codoni) o non-sinonime, che a loro volta possono essere suddivise in mutazioni *missenso* (in cui viene codificato un amminoacido differente) e in mutazioni *nonsense* (in cui la sequenza di DNA risulta in un prematuro codone di stop e viene così prodotta una proteina incompleta/tronca e solitamente non funzionale). Gli SNP invece presenti nelle regioni non codificanti possono alterare il livello di espressione di un gene, ed è stata dimostrata una correlazione tra SNP in regioni non codificanti e differenti patologie nell'umano (Zhang e Lupski, 2015) o in specie modello.

Il termine *variant calling* fa riferimento all'utilizzo del sequenziamento di nuova generazione (NGS) per analizzare le differenze nel genoma di un individuo o una popolazione di individui allo scopo di ottenere più informazioni possibile sulla variabilità genetica all'interno di una determinata specie. Una volta ottenute le sequenze, queste vengono allineate al genoma di riferimento e mappate, così da identificare i polimorfismi di ciascun individuo. In particolar modo il sequenziamento dell'intero genoma (WGS) viene utilizzato per ottenere le informazioni più complete, al fine di fornire un dataset utile a successivi studi funzionali o studi finalizzati ad individuare marcatori molecolari associati a differenze fenotipiche.

2. Scopo del lavoro

Dato l'importante valore economico dei cefali di sesso femminile, di gran lunga superiore a quello di sesso maschile, 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, anche prima della maturità sessuale, porterebbero enormi vantaggi all'industria del cefalo. Tale processo richiederebbe l'utilizzo di marcatori genetici legati al sesso affidabili, che non sempre è possibile identificare nei pesci ossei.

In un recente studio di Ferraresso et al. (2021) basato sull'analisi dell'intero genoma di *M. cephalus*, sono state identificate, su popolazioni di diversa origine geografica, due mutazioni missenso (denominate M1) nel gene codificante per il recettore dell'ormone follicolo-stimolante (*fshr*) presenti in eterozigosi (wt/m1) nella gran parte degli individui con fenotipo maschile. Un certo numero di maschi, definiti "non-conformi", hanno però mostrato un genotipo wt/wt, normalmente associato al fenotipo femminile. La frequenza di maschi non-conformi si è mostrata relativamente bassa (6-18%) nelle popolazioni italiane analizzate, mentre ha raggiunto una frequenza prossima al 50% in una popolazione dell'Egeo.

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 selezionati maschi conformi e non-conformi di diverse popolazioni è stato sequenziato a medio coverage (25X) e le varianti genetiche identificate sono state confrontate tra i due gruppi.

3. Materiali e metodi

3.1. Campioni utilizzati nello studio

I campioni di DNA utilizzati nel presente lavoro di tesi derivano da un precedente studio (Ferraresso et al. 2021). In tutto sono stati utilizzati 40 individui di sesso maschile, 20 maschi non-conformi (genotipo wt/wt) e 20 maschi conformi (genotipo wt/m1). I campioni sono stati scelti in modo da rappresentare equamente le due regioni geografiche di appartenenza Egeo (KAV) e Tirreno (TIR), per ciascuna area sono stati quindi selezionati 10 maschi conformi e 10 maschi non-conformi. Le caratteristiche genotipiche e di origine geografica di ciascun individuo sono riportate in Tabella 2.

Gruppo	Campione	Luogo di provenienza	Genotipo
TIR 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
	TOR_M189	Laguna di Tortolì (NU)	wt/m1
	TOR_M190	Laguna di Tortolì (NU)	wt/m1
TIR non-conformi	ORB_M19	Laguna di Orbetello (GR)	wt/wt
	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
KAV 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 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

Tabella 2. Caratteristiche genotipiche e di origine geografica di ciascun individuo

3.2. Preparazione delle librerie genomiche

I campioni di DNA selezionati sono stati sottoposti alla preparazione di librerie genomiche utilizzando il kit Illumina DNA Prep Tagmentation (cat#20018704). Come prima cosa è stata testata la fattibilità di utilizzo del kit Illumina con volumi di reagenti dimezzati rispetto a quanto indicato dal protocollo della ditta produttrice.

Per verificarne l'efficacia, 2 campioni (KAV_M02 e KAV_M03) sono stati processati sia con dosi di reagenti standard che dimezzate. Le librerie ottenute sono state comparate sia in termini di concentrazione che di dimensioni, prima con una quantificazione fluorimetrica al QuBit™ dsDNA High Sensitivity e in seguito con elettroforesi capillare tramite 2100 bioanalyzer High Sensitivity DNA Assay.

Per entrambi i campioni, la resa delle librerie è risultata essere comparabile tra i due protocolli. Per quanto riguarda le dimensioni, nonostante le dimensioni medie delle librerie fossero paragonabili tra i due protocolli, sono risultate essere più adatte al sequenziamento Illumina le librerie ottenute con il protocollo dimezzato che hanno mostrato una scarsa presenza di frammenti di lunghezza superiore a 1 kb (vedi Tabella 3). Si è quindi preferito usare, dove la concentrazione di DNA iniziale lo rendesse possibile, il protocollo modificato dimezzando la quantità di reagenti.

Campione	Qubit™ Concentrazione (ng/ µl)		Bioanalyzer Dimensioni medie (bp)		Bioanalyzer Range dimensioni	
	Protocollo Standard	Protocollo dimezzato	Protocollo Standard	Protocollo dimezzato	Protocollo Standard	Protocollo dimezzato
KAV_M02	9,46	7,05	515	527	400 - 2000	400 – 1000
KAV_M03	9,90	9,50	472	522	350 - 2000	400 – 1000

Tabella 3. Confronto tra protocollo Illumina DNA Prep Tagmentation standard e dimezzato.

Per 38 campioni sono state utilizzate le indicazioni della casa produttrice dimezzando le dosi dei reagenti ad eccezione di 2 maschi non-conformi CAB (CAB_M61 e CAB_M62) dove è stato utilizzato il protocollo con i volumi standard a causa della ridotta quantità di DNA di partenza (vedi sotto).

3.2.1. Protocollo Illumina DNA Prep Tagmentation

Per ciascuno dei 40 campioni inclusi nello studio, il protocollo di sintesi delle librerie è stato eseguito utilizzando 100 ng di DNA di partenza, ad eccezione di CAB_M61 (68,4 ng) e CAB_M83 (91,35 ng) per i quali la concentrazione non era sufficiente a raggiungere la quantità iniziale prevista. La sintesi di librerie genomiche tramite il protocollo DNA Prep Tagmentation prevede diverse fasi qui di seguito elencate.

3.2.1.1. Tagment Genomic DNA

La tagmentazione è una reazione mediata da trasposoni che combina la marcatura con la frammentazione del DNA in una singola e rapida reazione. Il primo passaggio prevede l'utilizzo di trasposoni legati alle microsfere (*Bead-Linked Transposomes* - BLT) per frammentare il DNA; l'utilizzo di BLT permette di eseguire una reazione di tagmentazione più uniforme rispetto alla medesima reazione in soluzione. Il trasposone è costituito da un enzima chiamato Tagmentasi (una transposasi) coniugato con opportuni oligonucleotidi che permette il simultaneo taglio del DNA e ligazione del frammento così ottenuto con gli adattatori per il sequenziamento.

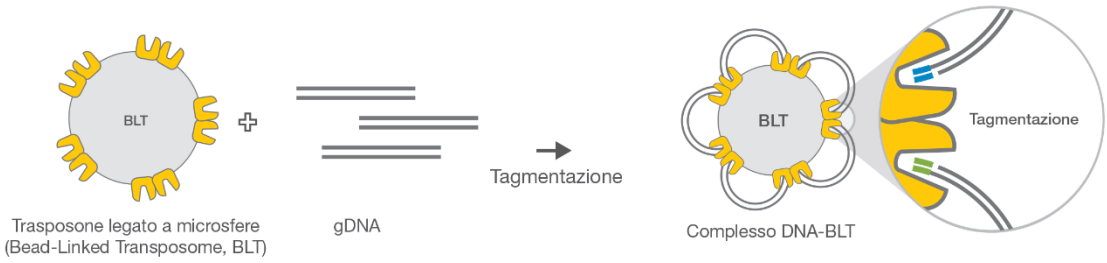

Figura 12. Chimica dei trasposoni legati alle microsfere Illumina. I trasposoni legati alle microsfere mediano contemporaneamente la frammentazione del gDNA e la ligazione dei primer di sequenziamento Illumina (www.Illumina.com)

La procedura si è articolata come segue:

Per ciascun campione, è stato prelevato un volume di DNA tale da ottenere una quantità di 100 ng. Nel caso il volume di DNA necessario fosse inferiore a 15 µl, al campione è stata aggiunta acqua *nuclease-free* per raggiungere il volume finale di 15 µl.

Per ciascun campione è stata, quindi, allestita una miscela di reazione, del volume finale di 10 µl, contenente:

BLT	5 µl
TB1 (*Tagmentation Buffer* 1)	5 µl

La miscela così composta è stata aggiunta a ciascun campione. I campioni sono stati quindi incubati a 55°C per 15 minuti.

3.2.1.2. Post Tagmentation Cleanup

Al termine dei 15 minuti, la reazione di tagmentazione deve essere interrotta per evitare una eccessiva frammentazione del DNA. A questo scopo, a ciascuna miscela di reazione sono stati aggiunti 5 µl di Tagment Stop Buffer (TSB) e in seguito i campioni sono stati incubati a 37°C per 15 minuti.

Al termine del programma, grazie all'aiuto di un supporto magnetico in grado di trattenere le microsfere BLT, queste sono state sottoposte a 3 lavaggi con Tagment Wash Buffer (TWB) allo scopo di eliminare i reagenti della reazione di tagmentazione e lavare il complesso "trasposone-DNA ligato" prima dell'amplificazione con la PCR.

3.2.1.3. Amplify Tagmented DNA

In questo passaggio DNA ligato viene amplificato utilizzando un programma PCR a ciclo limitato. Grazie alla PCR vengono aggiunti gli Index 1 (i7), gli Index 2 (i5), che permetteranno il pooling e il successivo demultiplexing delle diverse librerie dopo sequenziamento, e le sequenze necessarie per la generazione di cluster di sequenziamento.

Per ciascun campione è stata, quindi, allestita una miscela di reazione, del volume finale di 20 µl, contenente:

EPM (*Enhanced* PCR *Mix*)	10 µl
Acqua *nuclease-free*	10 µl

La miscela di reazione è stata aggiunta a ciascun campione, al quale sono stati poi addizionati 5 µl di index i7 e i5 pre-miscelati. È importante che ciascun campione venga indicizzato con index diversi per permettere il successivo pooling delle librerie.

Il profilo termico utilizzato è stato il seguente:

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

3.2.1.4. Clean Up Libraries

Le librerie così ottenute sono state purificate utilizzando le *Sample Purification Beads* (SPB). Questo passaggio, tramite l'utilizzo di specifiche quantità di biglie, permette di selezionare i frammenti di DNA sulla base della lunghezza e può essere quindi utilizzato per rimuovere i primers/dimeri presenti in soluzione nonché di rimuovere le molecole di DNA di dimensioni inferiori o superiori ad un determinato range.

Come prima cosa, i prodotti di PCR sono stati posizionati nel supporto magnetico allo scopo di separare il DNA dalle BLT, trasferendo 23 µl di prodotto di PCR in una nuova piastra al quale sono stati aggiunti 62 µl per raggiungere un volume totale di 85 µl.

Nella piastra contenente i prodotti di PCR sono stati poi aggiunti, per ciascun pozzetto, 45 µl di SPB. Dopo aver mescolato e incubato per 5' a temperatura ambiente su un agitatore fino ad ottenere una soluzione omogenea, la piastra è stata trasferita su supporto magnetico attendendo che le biglie si disponessero a formare un "anello" visibile sulla parete del pozzetto.

Per ciascun campione, 125 µl di surnatante sono stati quindi trasferiti su un nuovo pozzetto al quale sono stati aggiunti 15 µl di SPB. Dopo aver mescolato e incubato per 5' a temperatura ambiente su un agitatore fino ad ottenere una soluzione omogenea, la piastra è stata trasferita su supporto magnetico e il surnatante è stato questa volta eliminato, evitando di prelevare anche le biglie, legate al DNA delle dimensioni di interesse.

Continuando a mantenere la piastra sul supporto magnetico, sono stati effettuati 2 lavaggi con 200 µl di etanolo 80%. Una volta evaporato l'etanolo residuo, la piastra è stata rimossa dal supporto magnetico e il DNA è stato eluito dalle biglie con 25 µl di *Resuspension Buffer* (RSB).

3.2.2. Protocollo Qubit™ dsDNA High Sensitivity

La concentrazione delle librerie genomiche è stata misurata utilizzando un kit di quantificazione specifico per DNA (Qubit™ dsDNA High Sensitivity), utilizzando 2 µl per ogni campione. Tale metodica si basa sull'utilizzo di un fluorimetro (QuBit™) capace di quantificare accuratamente concentrazioni di DNA comprese tra 0,2 e 100 ng/µL. La soluzione di lavoro Qubit™ è stata preparata con 199 µl di Qubit™ buffer e 1 µl di fluoroforo Qubit™ (rapporto 1:200) per ogni campione da esaminare. In primo luogo

viene eseguita la calibrazione dello strumento mediante l'utilizzo di 2 standard contenenti concentrazioni note di DNA. A tale scopo sono stati utilizzati 190 µl di soluzione di lavoro 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 delle librerie, invece, sono stati utilizzati 198 µl di soluzione di lavoro e 2 µl del campione di DNA da analizzare. Le miscele così composte vengono brevemente agitate, centrifugate ed incubate a temperatura ambiente per almeno 2 minuti.

In seguito alla lettura, lo strumento restituisce valori espressi in ng/µl.

3.2.3. Valutazione qualitativa DNA

Per ciascuna libreria, la qualità e le dimensioni del campione sono state ricavate con Bioanalyzer High Sensitivity DNA Assay per 36 campioni e con High Sensitivity D5000 ScreenTape® per i rimanenti 4 campioni seguendo le indicazioni della casa produttrice.

3.2.3.1. Protocollo 2100 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 presenta 16 pozzetti, di cui 11 destinati ai campioni, collegati tra loro da microcanali di vetro contenenti 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 (per eseguire l'analisi di 11 campioni il tempo impiegato è di circa 45 minuti) e sensibilità aumentata. La rilevazione dei frammenti di DNA avviene attraverso una fluorescenza laser-indotta che sfrutta un colorante fluorescente intercalante la molecola di DNA.

Innanzitutto, è necessario permettere al *gel-dye mix*, stoccato a 4°C, di equilibrarsi a temperatura ambiente per 30 minuti prima dell'uso, al riparo dalla luce. Successivamente vengono posti 9 µL di *Gel-Dye* Mix nel fondo del pozzetto indicato con la lettera G cerchiata di nero (vedi Figura 13), prestando attenzione a non formare bolle d'aria, che potrebbero alterare il risultato.

Dopo essersi assicurati che lo stantuffo della siringa sia posizionato esattamente su 1 ml, viene chiusa la *chip printing station* e viene fatta pressione sullo stantuffo finché non sia correttamente tenuto da una molletta. Si attende esattamente un minuto e si rilascia lo stantuffo. Dopo essersi assicurati che lo stantuffo si sia mosso come minimo a 0,3 ml, si può lentamente sollevarlo fino a portarlo alla posizione iniziale di 1 ml. In questo modo il gel è stato distribuito in maniera omogenea all'interno dei micro canali.

Successivamente si caricano 9 µL di *Gel-Dye* Mix nei tre pozzetti indicati con la lettera G mentre nei rimanenti pozzetti vengono invece caricati 5 µL di *Marker.*

Successivamente, in un pozzetto dedicato (indicato con il simbolo di una scala) viene aggiunto 1 µL di *Ladder*, un marcatore di peso molecolare che permette di identificare la lunghezza dei frammenti contenuti nelle librerie. In ciascuno degli 11 pozzetti destinati ai campioni viene invece aggiunto 1 µL di DNA. Il Chip viene successivamente vortexato per 1 minuto a 2400 rpm e poi inserito nell'*Agilent 2100 Bioanalyzer* per la corsa e la relativa analisi.

Figura 13. Chip High Sensitivity DNA

Nonostante con il *Bioanalyzer* sia possibile eseguire anche una quantificazione del campione, calcolando l'area sottesa al picco corrispondente all'intensità di fluorescenza rilevata al passaggio dell'amplicone, è stato deciso di utilizzare il *Bioanalyzer* solo per valutare la qualità, e di utilizzare il Qubit™ per la quantificazione.

3.2.3.2. TapeStation 2200 e protocollo High Sensitivity D5000 ScreenTape®

La TapeStation 2200 effettua la separazione elettroforetica di acidi nucleici e proteine, il sistema High Sensitivity D5000 ScreenTape® contiene una matrice di gel per separare i campioni di acido nucleico in base al peso molecolare mediante elettroforesi ed è progettato per l'analisi di molecole di DNA da 100 a 5000 bp, con una maggiore accuratezza tra 400 e 5000bp. È possibile analizzare fino ad un massimo di 16 campioni contemporaneamente e, se non si utilizza la totalità delle 16 linee presenti, è permesso conservare a 2-8°C il *device* ScreenTape® utilizzato. Per questo studio si è preferito utilizzare questa metodica per analizzare i rimanenti 4 campioni in un'ottica di minor spreco rispetto al Bioanalyzer, in cui un chip deve essere completamente utilizzato.

Figura 14. D5000 ScreenTape®

Il ladder è stato preparato miscelando 2 µL di High Sensitivity D5000 Sample Buffer con 2 µL High Sensitivity D5000 Ladder in una *strip* di provette.

I campioni da analizzare sono stati preparati mescolando 2 µL di High Sensitivity D5000 Sample Buffer con 2 µL di campione di DNA nella stessa *strip* di provette.

Sono stati poi vortexati e centrifugati per 1 minuto a 2000 rpm. Infine, i campioni sono stati poi caricati nella TapeStation è si è avviata l'analisi.

3.2.4. Sequenziamento Illumina

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

Per il sequenziamento effettuato in questa tesi è stato deciso di utilizzare una *lane* (1/4) di *flowcell* NovaSeq S4, la quale fornisce fino a 3 TB di dati ed è ideale per sequenziamenti ad alta intensità. Una *lane* di *flowcell* NovaSeq S4, sequenziata in modalità 150 PE, produce in media 750 Gb di dati. Nel caso di 40 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 20X. Il sistema di sequenziamento utilizzato per la *flowcell* NovaSeq S4 è NovaSeq 6000, il quale utilizza il sequenziamento a due canali, che richiede solo due immagini per codificare i dati per quattro basi di DNA, un'immagine dal canale rosso e un'immagine dal canale verde. Una mancata identificazione viene indicata con una N. Le mancate identificazioni si verificano quando un cluster non attraversa il filtro, non viene eseguita la registrazione o un cluster si è spostato al di fuori dell'immagine. Le intensità di ciascun cluster sono estratte dalle immagini del canale rosso e del canale verde e sono confrontate tra di loro fornendo quattro popolazioni distinte, in cui ogni popolazione corrisponde a una base. Il processo di identificazione delle basi determina a quale popolazione appartiene ciascun *cluster*.

Figura 15. NovaSeq S4 flow (www.illumina.com)

Ad ogni base viene assegnato un punteggio qualitativo (*Q-score*), che corrisponde ad una previsione della probabilità di un'identificazione delle basi errata (www.illumina.com).

Punteggio qualitativo Q(X)	Probabilità di errore
Q40	0,0001 (1 su 10.000)
Q30	0,001 (1 su 1.000)
Q20	0,01 (1 su 100)
Q10	0,1 (1 su 10)

Tabella 4. Correlazione tra Punteggio qualitativo Q(x) e Probabilità di errore

3.2.5. Analisi risultati sequenziamento illumina attraverso Trim Galore!

Una volta concluso il sequenziamento, i dati di sequenza di ciascuna libreria vengono restituiti in formato FASTQ.

Un file FASTQ è un file di testo che contiene i dati della sequenza nucleotidica e la corrispettiva qualità per ogni nucleotide. Sia la lettera della sequenza che il punteggio di qualità sono codificati ciascuno con un singolo carattere ASCII ("*American Standard Code for Information Interchange*", un sistema di codifica dei caratteri a 8 bit).

Per una corsa a lettura singola, viene creato un file FASTQ di lettura 1 (R1) per ogni campione per *lane* della *flowcell*. Per una corsa *paired-end*, vengono creati un file FASTQ R1 e un file Read 2 (R2) per ogni campione per ciascuna *lane* (www.illumina.com).

Come prima cosa, la qualità delle sequenze è stata analizzata tramite il software fastQCv (ref). Questo software esegue una serie di analisi che permettono di determinare la qualità delle sequenze grezze in base a diversi parametri, i principali sono: i) *quality score* per ciascuna posizione delle sequenze, ii) *quality score medio* sull'intera lunghezza delle sequenze, iii) contenuto in basi N non determinate, iv) presenza di sequenze sovra rappresentate, v) livello di contenuto di adattatori.

Successivamente le sequenze grezze sono state sottoposte a *trimming*, un processo tramite il quale, se presenti, vengono eliminate porzioni di sequenze di bassa qualità o eventuali residui di adattatori. Per eseguire questo processo è stato utilizzato il software Trim Galore!.

Trim Galore! è uno script progettato dal Baraham Institute che permette di eseguire un *trimming* automatico delle sequenze ottenute attraverso il sequenziamento Illumina, grazie al riconoscimento dei primi 12 nucleotidi degli adattatori nextera ('CTGTCTCTTATA'). I comandi eseguiti sono stati i seguenti:

```
for s in $(cat Sample_list.txt);

do
```

```
trim_galore -q 25 --length 70 --paired --clip_R1 5 --clip_R2 5 --nextera
--cores 6 -o ./Trimmed --basename ${s} ./RawSeqs/${s}_R1.fastq.gz
./RawSeqs/${s}_R2.fastq.gz

done
```

--q 25

Per tagliare le estremità di bassa qualità dalle *reads* oltre alla rimozione dell'adattatore. La qualità minima accettabile è stata impostata su un Phred score di 25, che corrispondono ad una probabilità pari a 0,00316 che quella base venga letta in modo errato.

--length 70

Per eliminare le *reads* di lunghezza inferiore a 70 bp a causa del trimming dovuto alla qualità o al taglio dell'adattatore.

--paired

Questa opzione esegue il trimming per i file accoppiati. Per superare il test di convalida, entrambe le sequenze di una coppia di sequenze devono avere una certa lunghezza minima che è regolata dall'opzione –length.

--clip_R1 5 e --clip_R2 5

Per rimuovere dall'estremità 5' le prime 5 basi sia per forward (R1) che reverse (R2), in quanto si è cercato di rimuovere quel rumore di fondo sempre presente all'inizio di una lettura.

--nextera

Per far riconoscere la corretta sequenza dell'adattatore da tagliare.

--cores 6

Per aumentare il numero di *core* utilizzabili e rendere più veloce il processo.

-o /home/sferraresso/elab/Cefalo_2022/Trimmed

Per indicare il percorso della cartella su cui salvare i file di output.

--basename ${s}

Per indicare il prefisso del nome del file di output.

3.2.6. Mappatura delle sequenze sul genoma attraverso BWA-MEM

Per ciascun campione, i file di sequenza sono stati quindi mappati sul genoma di riferimento di *Mugil cephalus* tramite il software BWA.

BWA è un software per la mappatura di sequenze in relazione ad un genoma di riferimento. Esistono 3 differenti algoritmi, BWA-backtrack, BWA-SW e BWA-MEM. BWA-backtrack è utilizzato per sequenze Illumina fino a 100bp, i rimanenti possono essere utilizzati per sequenze più lunghe, da 70bp a 1 Mbp. BWA-MEM e BWA-SW condividono funzionalità simili ma BWA-MEM, l'ultimo ad essere stato rilasciato, è generalmente consigliato per l'elaborazione di dati di alta qualità in quanto è più veloce e più accurato (http://www.bio-bwa.sourceforge.net/).

L'algoritmo funziona allineando i segmenti con le massime corrispondenze esatte (*maximal exact matches* - MEMs) confrontandole con il genoma di riferimento.

I comandi eseguiti sono stati i seguenti:

```
mkdir ./BWA_mem

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
```

-t 15

Per aumentare il numero di cores utilizzati dal software e rendere più veloce il processo.

-c 1

Per scartare le sequenze che vengono mappate in più punti del genoma.

-M

Per contrassegnare gli *split hit* più brevi come secondari (per compatibilità con Picard).

samtools sort -@15

Per utilizzare più *thread* e rendere più veloce il processo.

```
-o ./BWA/${s}.bam
```

Per indicare il nome del file di output.

Alla fine del processo sono stati ottenuti 40 file BAM, contenenti una rappresentazione binaria dei segmenti che sono stati allineati nella mappa genomica.

3.2.7. Creazione dei *read group* e SNP calling

Per poter rimuovere i duplicati presenti in un singolo segmento, si necessita di usare lo strumento 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), il quale necessita che ai file BAM sia assegnato un *read group*, così da differenziare i diversi campioni.

È stato quindi assegnato un differente *read group* per ogni campione con il seguente comando

```
#READ GROUP

for s in $(cat 11_20.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
```

Strumento che permette di sostituire i *read group* in un file BAM. In questo modo è possibile sostituire tutti i *read group* nel file di *input* con un unico nuovo *read group* e assegnare tutte le *read* a quel *read group* nel file BAM di *output*.

```
INPUT= ./BWA/${s}.bam
```

Per indicare il percorso del file di input.

```
OUTPUT= ./BWA/${s}_RG.bam
```

Per indicare il percorso del file di output.

```
RGID=${s}
```

Per assegnare un ID al *read group.*

RGLB=DNA

Per indicare che nella libreria di riferimento sono presenti sequenze di DNA.

RGPL=Illumina

Per indicare che la libreria è stata creata con la piattaforma “Illumina”.

RGPU=01

Stringa arbitraria indicante lo strumento utilizzato

RGSM=${s}

Per indicare il nome dei campioni a cui verrà assegnato il *read group*.

In seguito, è stato eseguito il comando *MarkDuplicates* di Picard per marcare le letture duplicate, risultato di un bias di PCR. Lo strumento “*MarkDuplicates*” confronta le sequenze nelle posizioni 5’ sia delle *reads* che delle *reads-paired* in un file SAM o BAM.

Per eliminare le letture duplicate, si è eseguito il comando

```
#REMOVE DUPLICATES

for s in $(cat 11_20.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
```

INPUT= ./BWA/${s}_RG.bam

Per indicare il percorso del file di input.

OUTPUT= ./BWA/${s}_RG_MD.bam

Per indicare il percorso del file di output.

M=./BWA/${s}_metrics.txt

Per creare un file di output con la statistica descrittiva di ogni campione

```
REMOVE_DUPLICATES = true
```

Per non scrivere i duplicati nel file di output.

```
CREATE_INDEX=true
```

Per creare un indice quando viene scritto il file BAM ordinato secondo le coordinate.

Infine, per ottenere un file VCF compresso il quale all'interno contenga le coordinate dei SNP identificati si è utilizzato lo strumento Bcf tools mpileup.

Un file VCF è un file di testo in cui sono contenute le informazioni riguardanti i SNP riscontrati nei campioni analizzati, tra cui il cromosoma (o *contig*) e 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.

È stato analizzato e fatto *SNP calling* sul *contig* 2337, il quale presenta al suo interno la regione corrispondente al gene *fshr* grazie al comando

```
#SNP CALLING

mkdir ./VCF

bcftools mpileup -b ./BAM_list.txt --fasta-ref
./Genome/Mcephalus_assembly.fa -q 20 -O z -o ./VCF/Cefalo_mpileup.vcf
--threads 18
```

```
-b ./BAM_list.txt --fasta-ref ./Genome/Mcephalus_assembly.fa
```

Per inserire come input i file BAM dei campioni analizzati e il genoma di *Mugil cephalus* presente in database in formato FASTA.

```
-q 20
```

Per assicurarsi che la qualità di mappatura minima utilizzabile per un allineamento sia di 20 Phred.

```
-O z
```

Per indicare il formato del file di output, "z" corrisponde ad un file VCF compresso.

```
-o ./VCF/Cefalo_mpileup.vcf
```

Per indicare il nome del file di output.

```
--threads 18
```

Per aumentare il numero di cores utilizzati dal software e rendere più veloce il processo.

Una rappresentazione schematica dell'intero processo di analisi bioinformatiche utilizzato è visualizzabile nella Figura 16.

Figura 16. Schematizzazione del workflow utilizzato per le analisi bioinformatiche

3.2.8. Calcolo F_{st}

Si è successivamente calcolato l'F_{st} sulle popolazioni del Tirreno e dell'Egeo, per osservare la differenza tra le frequenze alleliche presente tra conformi e non-conformi.

F_{ST} è la proporzione della varianza genetica dovuta alle differenze di frequenza degli alleli tra le popolazioni. F_{ST} ha valori che vanno da 0 a 1, 0 indica una situazione di completa panmissia nella quale non c'è alcuna differenza tra due popolazioni mentre un valore di 1 indica una completa separazione (Binelli e Ghisotti, 2017).

Un F_{st} elevato implica un notevole grado di differenziazione tra le popolazioni (Holsinger e Weir, 2009). In quest'ottica si possono definire bassi valori di F_{ST} minori di 0,1, valori medi quelli compresi tra 0,1 e 0,15-0,2, valori alti quelli superiori a 0,2 (Binelli e Ghisotti, 2017).

F_{st} è stato calcolato grazie al pacchetto adegenet con il linguaggio di programmazione R.

4. Risultati

4.1. Qualità delle librerie genomiche

L'analisi della resa e delle dimensioni di ciascuna libreria genomica, effettuate rispettivamente con metodo fluorimetrico QuBit™ ed elettroforesi capillare tramite Bioanalyzer2100 o Screen Tape, hanno permesso di evidenziare la buona qualità di tutti i campioni. I risultati di Bioanalyzer e Screen tape hanno dimostrato una dimensione media di 583,4 bp, con profili molto simili tra i differenti campioni analizzati.

I risultati del QuBit™ hanno dimostrato una concentrazione media di 5,8 ng/µl. (*range* 2,3-9,76 ng/µl), corrispondente a una molarità 15 nM (*range* 5,98-36,30 nM), più che sufficiente per soddisfare le richieste del centro di sequenziamento (2-30nM).

Campione	**Concentrazione (ng/µl)**	**Dimensione media (bp)**	**Molarità (nM)**
KAV_M02	7,05	522	13,86
KAV_M03	9,5	412	15,56
KAV_M04	7,88	671	12,85
KAV_M11	8,65	656	13,25
KAV_M13	8,98	652	15,68
KAV_M15	4,79	652	9,11
KAV_M16	6,4	606	15,00
KAV_M18	6,06	623	13,80
KAV_M19	2,77	587	7,76
KAV_M20	6,91	587	16,17
KAV_M01	4,52	637	9,37
KAV_M05	9,23	663	14,08
KAV_M07	8,33	677	12,51
KAV_M08	2,41	613	6,51
KAV_M09	5,47	552	17,18
KAV_M10	6,68	606	15,97
KAV_M12	3,95	564	12,85
KAV_M14	9,76	584	20,53
KAV_M17	2,3	568	7,28
KAV_M21	7,48	600	22,18

ORB_M19	8,21	548	26,89
ORB_M21	6,16	592	16,24
CAB_M61	2,35	647	5,98
CAB_M62	7,28	491	36,30
CAB_M63	4,37	557	13,31
CAB_M80	3,79	546	11,99
CAB_M83	3,09	547	10,78
TOR_M175	6,21	558	18,05
TOR_M195	5,45	598	13,50
TOR_M196	4	577	14,73
ORB_M04	6,39	570	16,10
ORB_M06	6,06	587	15,97
CAB_M03	5,9	551	15,43
CAB_M05	5,03	587	12,97
CAB_M06	5,12	575	14,48
CAB_M33	3,78	556	11,17
CAB_M34	5,63	603	13,90
TOR_M172	4,75	585	13,53
TOR_M173	5,85	531	23,00
TOR_M174	4,08	498	18,30

Tabella 5. Concentrazione e dimensione media dei campioni analizzati.

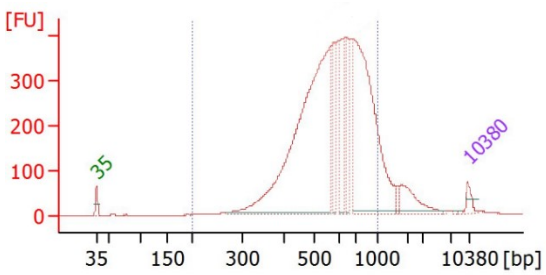

Figura 17. Rappresentazione grafica delle diverse dimensioni e della rispettiva quantità dei frammenti nel campione KAV_M21. Si può notare che la maggior parte dei frammenti è concentrata tra le 400 e 1000 basi.

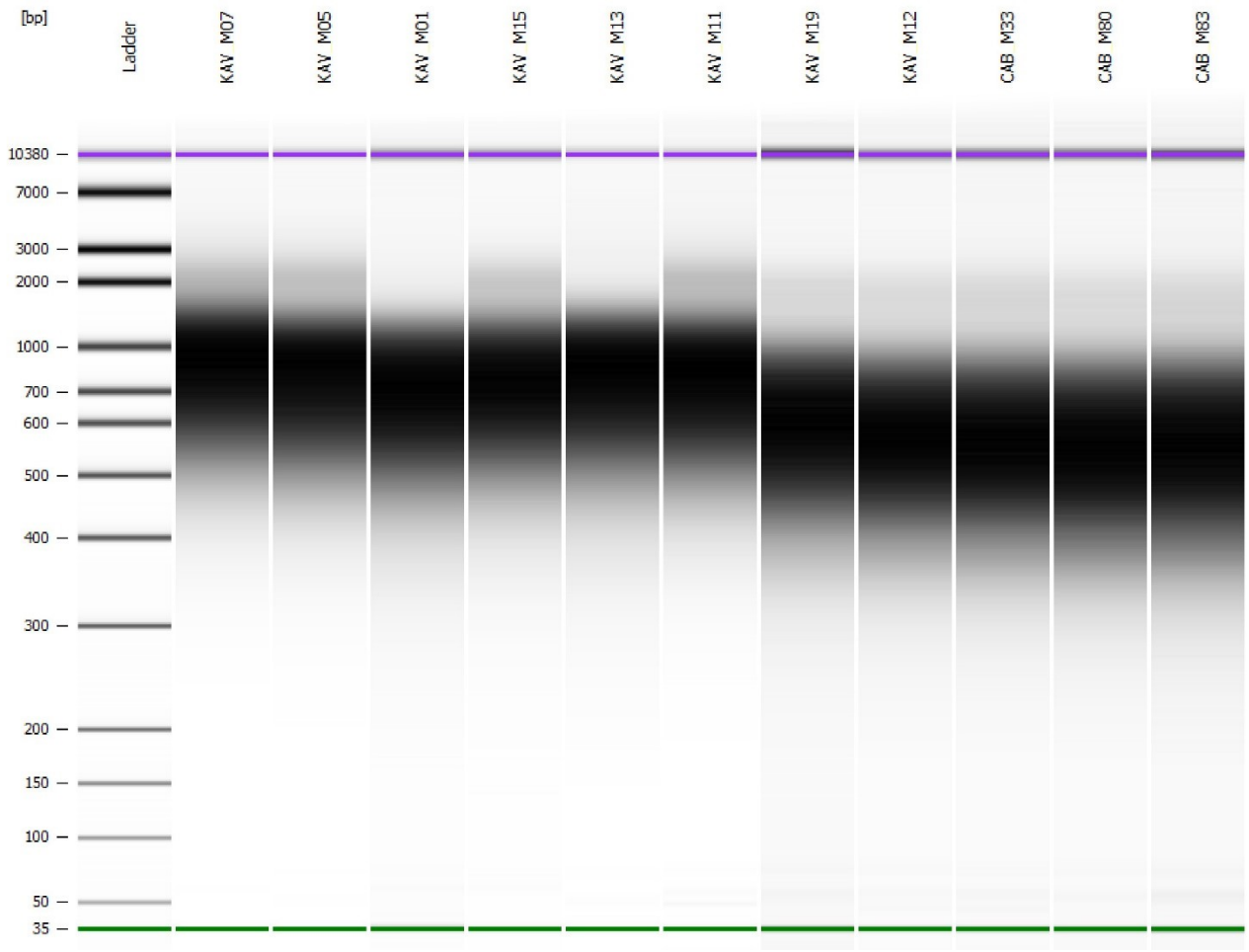

Figura 18. Esempio di risultati del Bioanalyzer 2100 High Sensitivity DNA Assay. Nell'asse verticale sono presenti le bp, nell'asse orizzontale è presente il ladder e i campioni analizzati nella corsa. Il ladder, presente nella prima colonna, è un marcatore di peso molecolare ed è necessario per stimare le dimensioni dei frammenti di interesse. Nelle colonne successive sono presenti alcuni dei campioni analizzati. Ad un colore più scuro corrisponde una maggiore quantità di frammenti di quella data dimensione.

4.2. *Quality Control* dei dati grezzi di sequenziamento

Il sequenziamento Illumina delle 40 librerie genomiche ha prodotto un totale di 5,89 miliardi di *read* con un minimo di 56,7 milioni ad un massimo di 104,6 milioni di *read* per campione (Figura 19), e una media di 73,6 milioni di *read* per campione, corrispondenti a 22 Gb sequenziate per ogni campione. La stima del *coverage* ottenuto per ciascun campione è stata quindi eseguita come segue:

$$Coverage = \frac{number\ of\ reads\ \times read\ base\ pairs\ \times 2}{Genome\ base\ pairs}$$

Il genoma di *Mugil cephalus* è stimato essere 0,8 Gb (equivalenti a 0,8 x 10^9 bp) (Raymond, et al., 2022). Per ogni campione quindi il *coverage* è risultato essere in media di 27X (da un minimo di 21X ad un massimo di 39X, vedi Figura 20), potendo così classificare il sequenziamento come medio *coverage*.

Per ciascun campione, la qualità delle sequenze grezze ottenute dal centro di sequenziamento è stata valutata con il software fastQC, il quale restituisce gli output grafici rappresentati in Figura 21 e Figura 22. Mediante questa rappresentazione grafica è possibile dunque definire la qualità delle sequenze per ciascuna posizione nucleotidica (Figura 21) oppure la distribuzione della qualità media delle sequenze nel loro complesso (Figura 22). Tutti i campioni hanno mostrato una qualità elevata delle sequenze grezze, con Phred score medi sempre superiori a 28, soglia che delimita le sequenze di qualità elevata dalle sequenze di qualità accettabile. Nessun campione ha mostrato, in alcuna posizione, un Phred score inferiore a 20, indice di sequenza di bassa qualità.

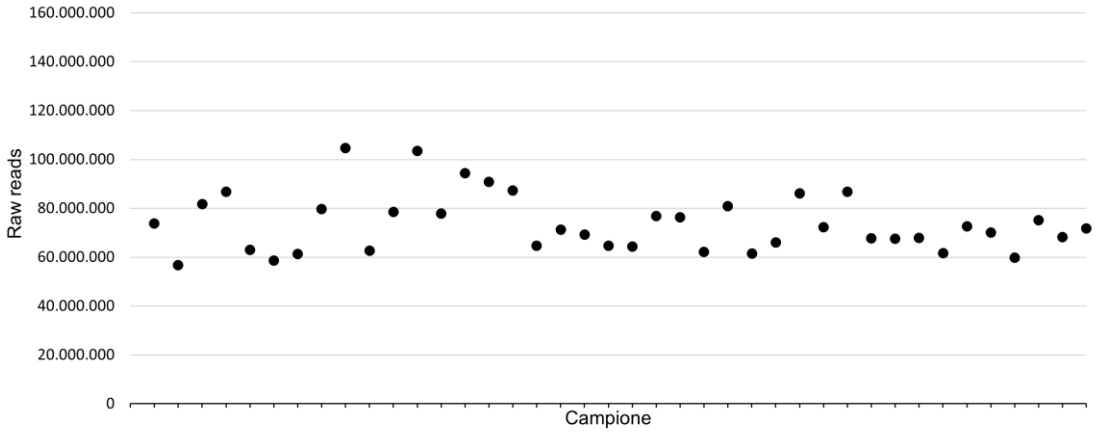

Figura 19. Distribuzione delle raw reads dei differenti campioni analizzati

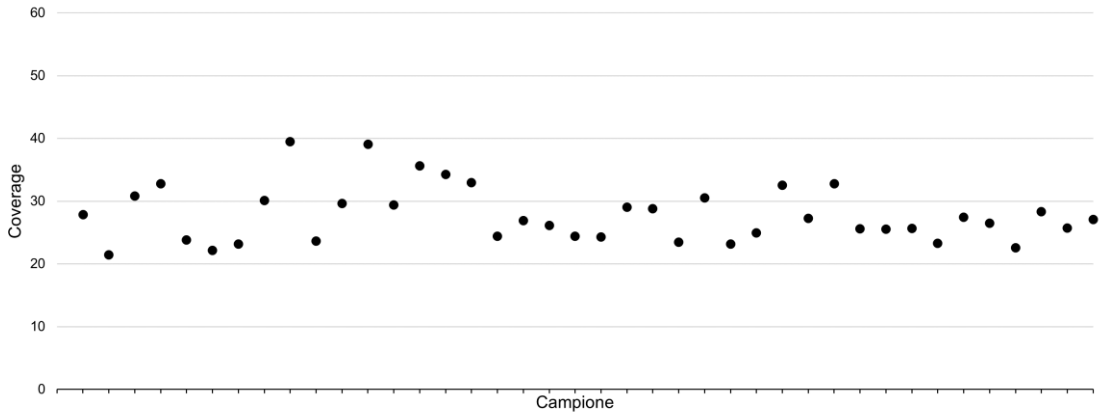

Figura 20. Coverage stimato per ogni campione

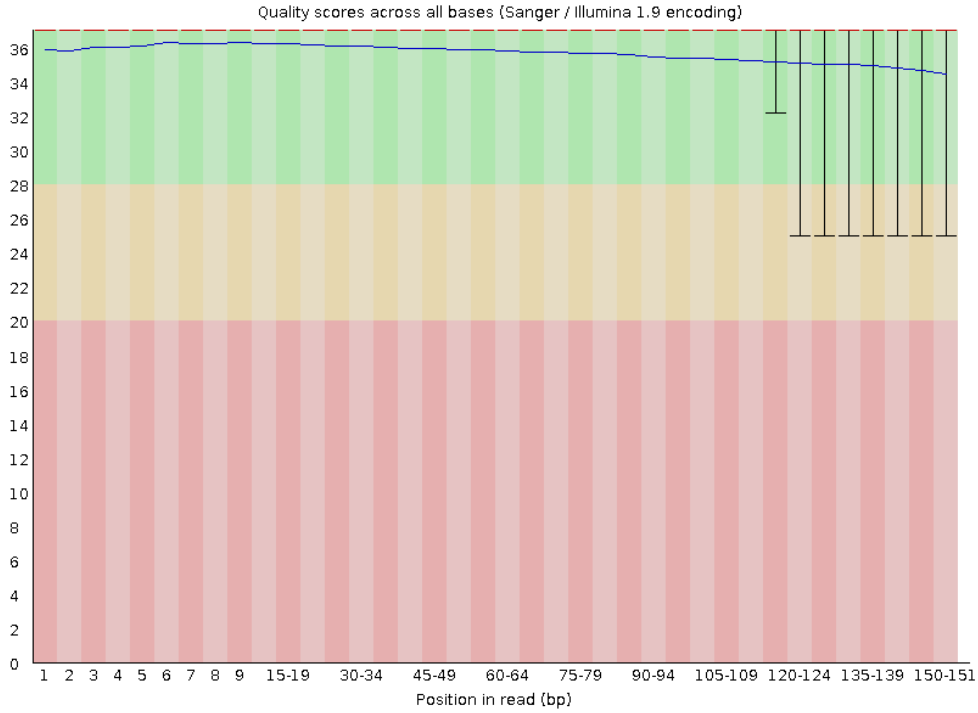

Figura 21. Esempio di grafico per permettere la visualizzazione della qualità, in termini di Phred score, delle reads del campione KAV_M07, creato grazie a FastQC. L'asse x indica la posizione della base nella lettura e l'asse y mostra i quality scores. Tutte le letture sono lunghe 151 bp. Per ogni posizione, c'è un diagramma box-and-whisker che mostra la distribuzione dei quality scores per tutte le reads in quella posizione. La linea rossa orizzontale indica il punteggio di qualità mediano, la linea blu indica la media. Le barre mostrano l'intervallo assoluto, che copre i valori più bassi (0° quartile) e più alti (4° quartile). Lo sfondo del grafico è codificato a colori per identificare i punteggi di qualità elevata (verde), accettabile (giallo) e bassa (rosso). Per ogni posizione in questo campione i valori di qualità non scendono mai al di sotto di 25, la qualità media è sempre superiore a 34 e la qualità di quasi tutte le reads è maggiore di 28, dimostrando una buona qualità del sequenziamento.

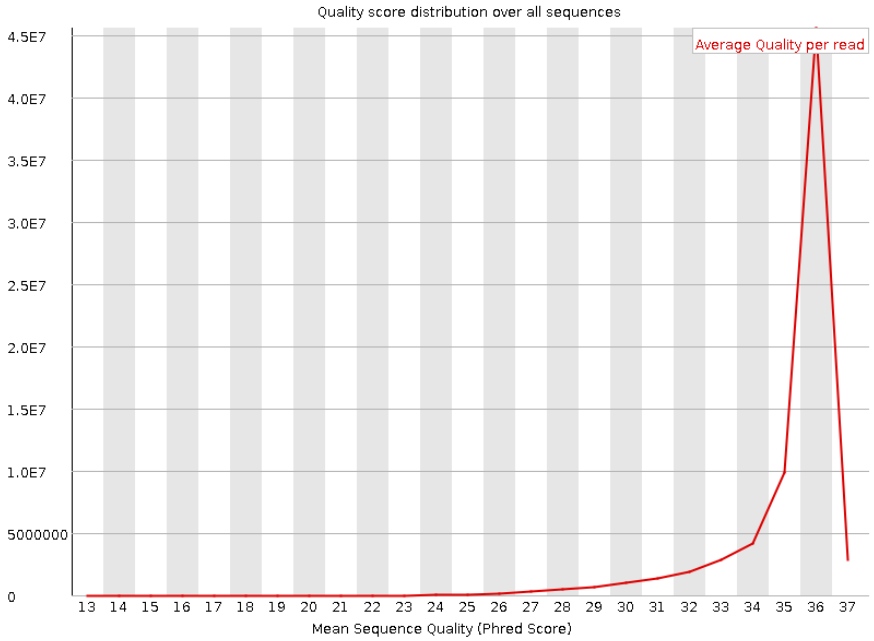

Figura 22. Esempio di grafico generato dai risultati di KAV_M07 calcolando la qualità media di ciascuna read e quindi tracciandone la distribuzione. Come evidenziato dal grafico, la quasi totalità delle read ha una Phred score superiore a 28 con un evidente picco (corrispondente al 90% delle read) a 36 Phred, permettendo di classificare questa lettura di ottima qualità.

4.3. Trimming delle sequenze grezze

Le sequenze grezze sono state quindi sottoposte a *trimming* tramite il software Trim Galore! allo scopo di eliminare l'eventuale presenza di adattatori che comprometterebbero la successiva fase di *mapping* nel genoma di riferimento. Il *trimming* è stato impiegato anche per eliminare eventuali nucleotidi di bassa qualità dalle sequenze, anche questo per evitare il più possibile errori nel sequenziamento che sarebbero potuti risultare in erronee attribuzioni di varianti nucleotidiche.

Anche il *trimming* ha confermato la buona qualità di sequenziamento, con una percentuale di *read* mantenute dopo questo passaggio superiore al 96% in tutti i campioni (vedi Figura 23 e Appendice A). Tramite FastQC è stata verificata la qualità dei campioni trimmati, nonostante la qualità delle sequenze grezze fosse già elevata (Figura 21) è stato possibile apprezzare un ulteriore miglioramento della qualità media soprattutto nelle ultime posizioni (un esempio è riportato in Figura 24).

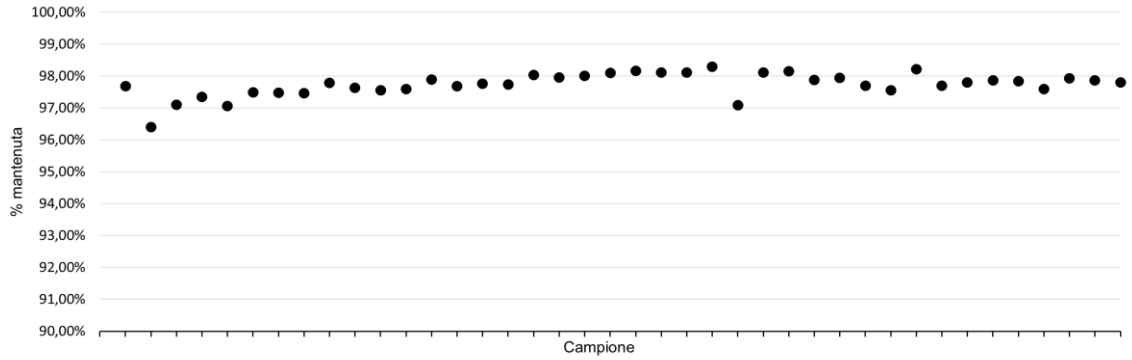

Figura 23. Distribuzione della percentuale di reads mantenute dopo il trimming dei differenti campioni analizzati

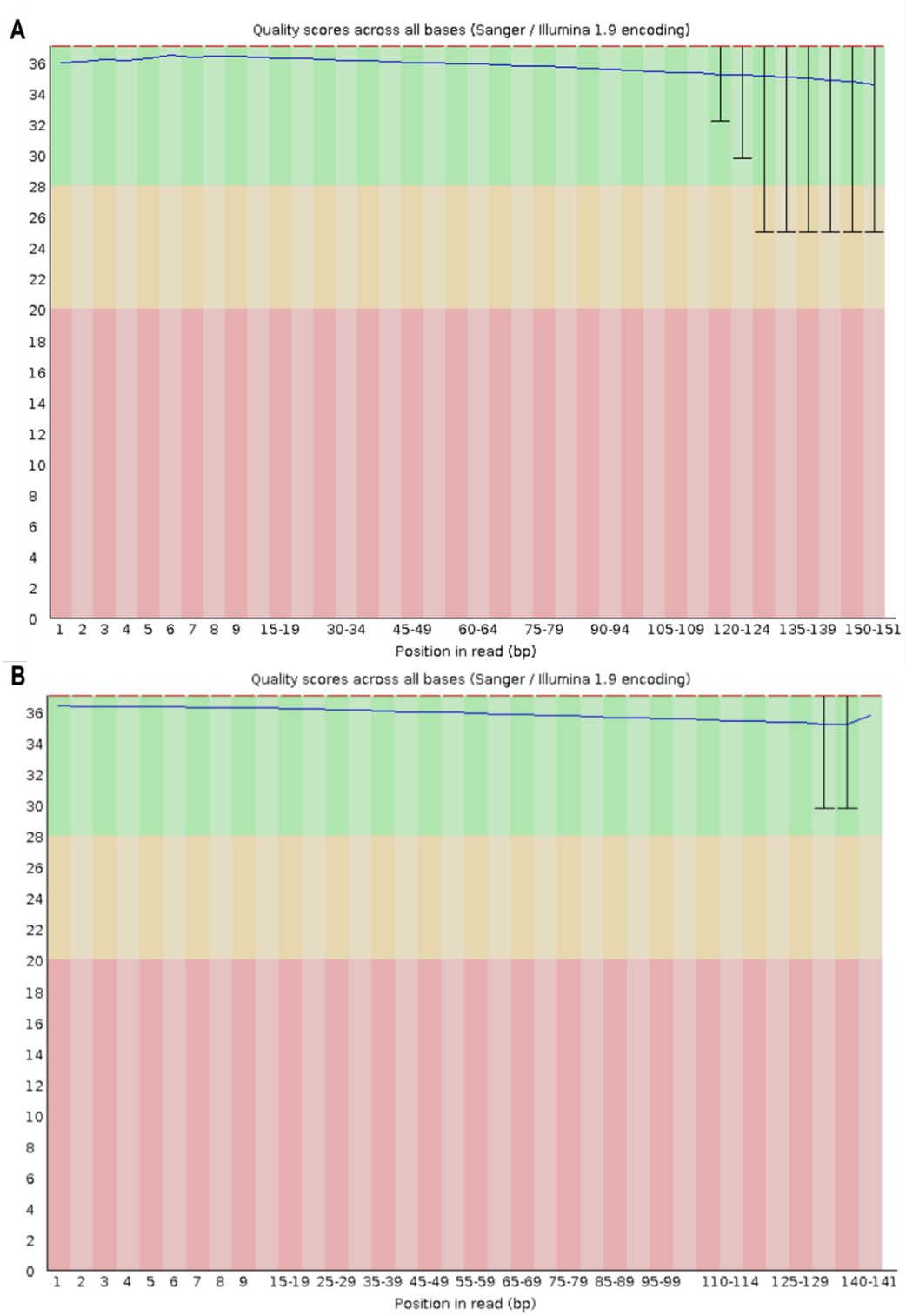

Figura 24. Esempio di grafico per permettere la visualizzazione della qualità, in termini di Phred score, delle reads del campione TOR_M196, creato grazie a FastQC. L'asse x indica la posizione della base nella lettura e l'asse y mostra i quality scores. Per ogni posizione, c'è un diagramma box-and-whisker che mostra la distribuzione dei quality scores per tutte le reads in quella posizione. La linea rossa orizzontale indica il punteggio di qualità mediano, la linea blu indica la media. Le barre mostrano l'intervallo assoluto, che copre i valori più bassi (0° quartile) e più alti (4° quartile). Lo sfondo del grafico è codificato a colori per identificare i punteggi di qualità elevata (verde), accettabile (giallo) e bassa (rosso). Nel grafico A è rappresentata la qualità del campione prima del trimming, nel grafico B dopo il trimming. Si può notare che dopo il trimming i valori di qualità non scendono mai al di sotto di 30, la qualità media è sempre superiore a 34 ed è stato possibile apprezzare un miglioramento della qualità media soprattutto nelle ultime posizioni.

4.4. Mapping sul genoma di riferimento e Variant calling

Le *read* di ciascun campione sono state allineate al genoma di *Mugil cephalus* (*mapping*) tramite BWA. Come riportato in materiali e metodi, allo scopo di rendere il più possibile affidabile la successiva fase di individuazione delle varianti nucleotidiche, sono state eliminate tutte le *read* che mappano in modo non univoco nel genoma. A seguito del *mapping,* una media di 99,14% *reads* sono state correttamente mappate sul genoma e una media del 97,87% sono state propriamente allineate in *pair*, ossia sia la sequenza *forward* che la *reverse* sono risultate allineate sulla stessa regione genomica.

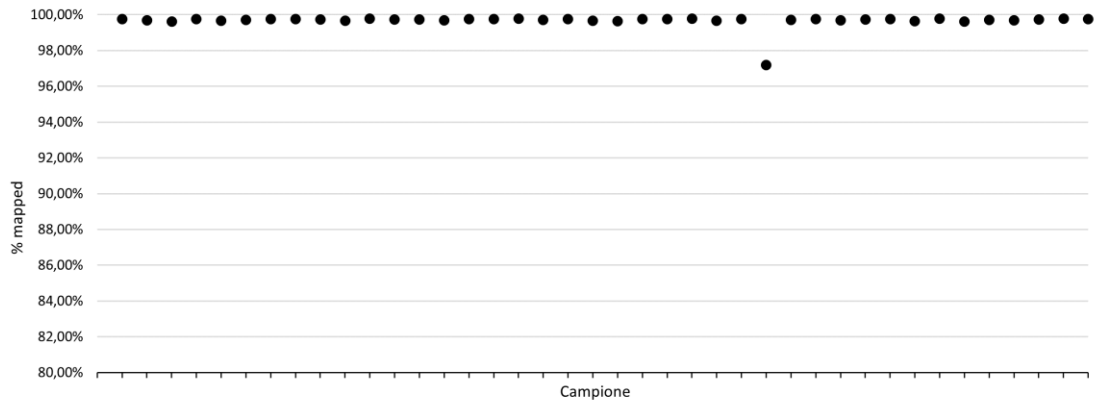

Figura 25. Distribuzione della percentuale di mappatura per ogni campione analizzato

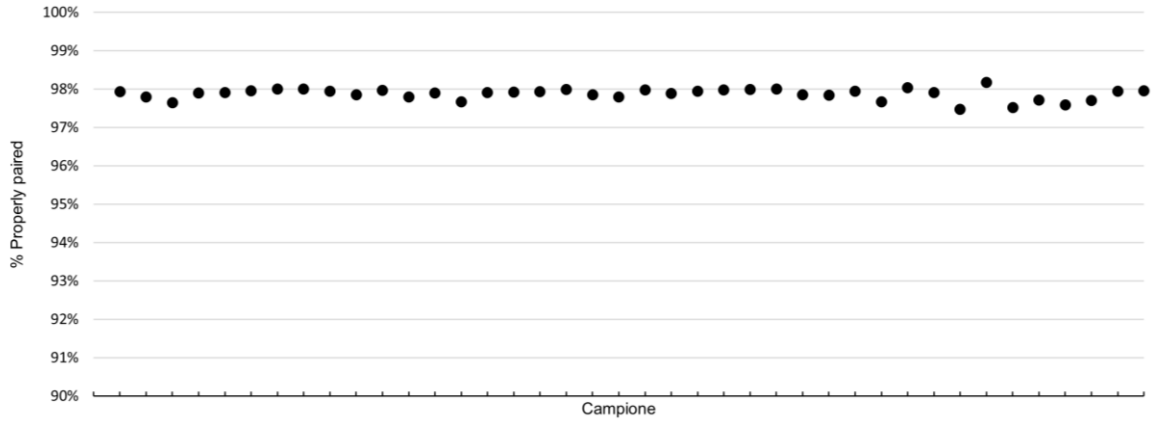

Figura 26. Distribuzione della percentuale di reads propriamente allineate in pair per ogni campione analizzato

La fase di identificazione delle varianti nucleotidiche (*Variant calling*) si è concentrata nel contig che contiene il gene *fshr*, il contig_2337 di 605 kb. In seguito al *variant calling* sono stati identificati in totale 22945 SNP, 4522 *Indels*, 100 siti (indel + SNP) multiallelici.

Con il termine *indel* si intende una mutazione rispetto al genoma di riferimento che può essere o una inserzione (***in**sertion*) o una delezione (***del**etion*). Generalmente *indel* lunghe 1-5 basi consecutive sono causate da un errore di trascrizione del genoma.

L'allineamento sul genoma di riferimento delle reads nelle regioni contigue alle *indel* non è accurato e presenta spesso errori di allineamento nell'intorno dell'*indel* che risultano nella identificazione erronea di varianti. Per questo motivo si è scelto di filtrare gli SNP posizionati in una finestra di ±5 nucleotidi dalle *indel* identificate, così come le *indel* stesse.

Nei siti mulitallelici sono stati ritrovati, nei diversi individui, più di una base differente dal genoma di riferimento. Si è deciso di eliminare anche i siti multiallelici, in quanto non pertinenti con la ricerca in esame.

Dopo aver rimosso le varianti multialleliche, le *indel* e gli SNP in loro prossimità sono risultati analizzabili 20699 SNP.

	VCF grezzo	**Rimozione SNP ± 5 bp da *indel***	**Rimozione *indel* e siti multiallelici**
Numero di SNP	22945	20708	20699
Numero di *Indels*	4522	4522	0
Numero di siti multiallelici	100	96	0
Numero di siti di SNP multiallelici	13	9	0

Tabella 6. Resoconto dei SNP mantenuti dopo ogni passaggio

4.1. Identificazione di mutazioni legate alla determinazione del sesso

L'analisi delle varianti genetiche legate alla determinazione del sesso si è focalizzata sugli SNP localizzati nel gene *fshr* e nella regione genomica immediatamente contigua (5Kb a monte e a valle del gene). La regione selezionata (denominata FSHR ± 5Kb) è risultata quindi essere di quasi 22 Kb (dalla posizione 369688 alla 391540 del contig_2337) nella quale sono stati identificati un totale di 676 SNP.

Le 3 varianti precedentemente identificate da Ferraresso et al. (2021) sono state utilizzate per confermare il gruppo di appartenenza degli individui inclusi nel presente lavoro di tesi. L'attribuzione è risultata corretta, confermando per 39 individui il genotipo stabilito in partenza. L'unica eccezione è risultato essere il campione KAV_M17, il quale era stato identificato come non-conforme in seguito a sequenziamento Sanger (genotipo wt/wt) mentre nel presente lavoro è risultato essere conforme (genotipo wt/m1). Non potendo escludere un errore di campionamento durante la sintesi delle librerie è stato deciso di tralasciare il campione KAV_M17 dalle successive analisi.

I genotipi di tutti i 676 SNP localizzati nella regione FSHR ± 5 Kb sono stati utilizzati in un'analisi multivariata per valutare la diversità genetica tra i gruppi di individui analizzati: i) maschi Tirreno conformi (TIR_c), ii) maschi Tirreno non-conformi (TIR_nc), iii) maschi

Egeo conformi (KAV_c) e iv) maschi Egeo non-conformi (KAV_nc). A tale scopo è stata eseguita l'analisi DAPC (*Discriminant analysis of principal components*).

La DAPC è progettata per identificare e descrivere gruppi di individui geneticamente correlati fornendo una valutazione visiva della differenziazione tra le popolazioni e il contributo dei singoli alleli alla struttura di popolazione. DAPC si basa su PCA (*Principal component analysis*) e DA (*Discriminant Analysis*). DAPC prevede la trasformazione dei dati utilizzando la PCA come passaggio precedente alla DA, il che garantisce che le variabili presentate nella DA non siano perfettamente correlate e che il loro numero sia inferiore a quello degli individui analizzati (Jombart, et al., 2010).

La PCA viene utilizzata per trovare un sistema di riferimento in grado di massimizzare la varianza delle variabili rappresentate, calcolando il peso da attribuire ad ogni variabile di partenza per poterle concentrare in una o più nuove variabili (dette componenti principali) che saranno una combinazione lineare delle variabili di partenza. La DA viene utilizzata per massimizzare la separazione tra i gruppi riducendo al minimo la variazione all'interno del gruppo (Jombart, et al., 2010).

Questa analisi ha permesso di evidenziare una chiara separazione tra maschi conformi e maschi non-conformi lungo la prima componente (vedi Figura 27), confermando la presenza nella regione FSHR ± 5 Kb di polimorfismi in grado di differenziare in modo accurato i due gruppi di individui. Lungo la seconda componente è visibile anche una separazione legata all'origine geografica dei campioni, separando i campioni Egeo dai campioni Tirreno.

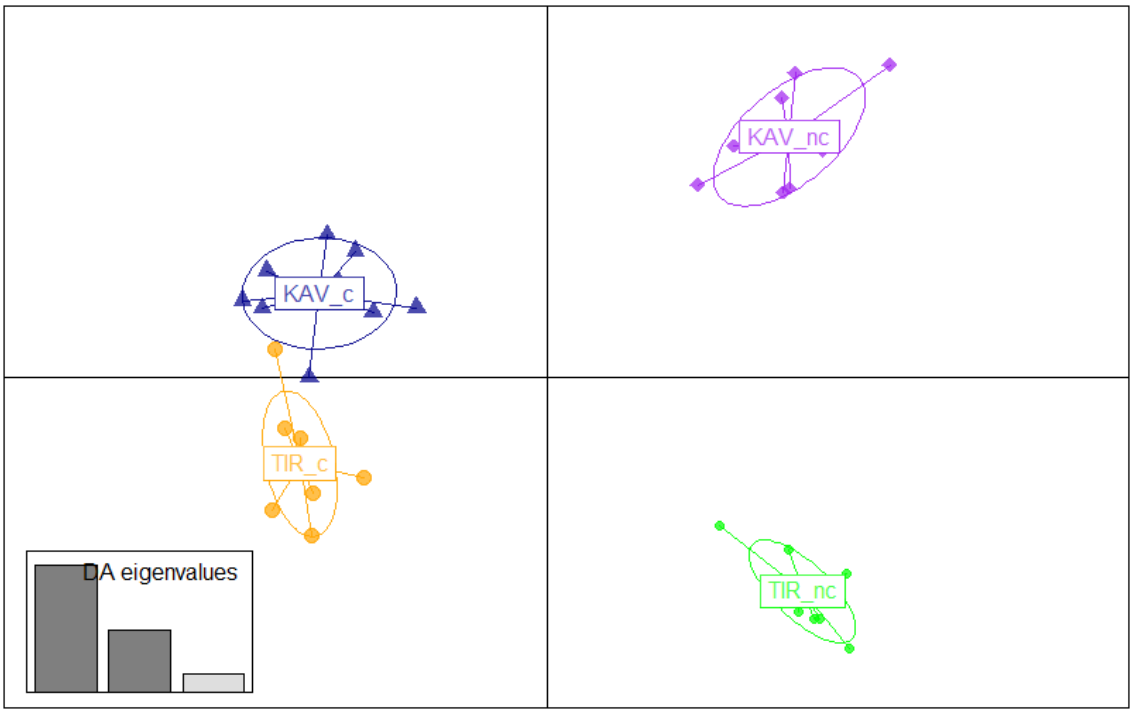

Figura 27. Discriminant analysis of principal components (DAPC) delle quattro differenti popolazioni analizzate. KAV_c: maschi conformi Egeo;KAV_nc: maschi non-conformi Egeo; TIR_c: maschi conformi del Tirreno e TIR_nc maschi non-conformi del Tirreno. Gli Eigenvalues raffigurati misurano la varianza delle componenti principali.

L'F_{ST} calcolato per ciascuna popolazione ha inoltre permesso di individuare le varianti la cui frequenza allelica risulta maggiormente sbilanciata nel confronto maschi conformi e maschi non-conformi. In Figura 28 è possibile visualizzare la distribuzione della varianza presente tra le due differenti popolazioni tra individui conformi e non-conformi. Per entrambe le popolazioni, si può notare una maggiore densità di SNP con F_{ST} più elevato a livello dell'ultima porzione del gene dell'*fshr*, indicando una maggiore differenziazione a livello di quelle posizioni tra i gruppi di maschi conformi e non-conformi, sia nel caso di Egeo che Tirreno.

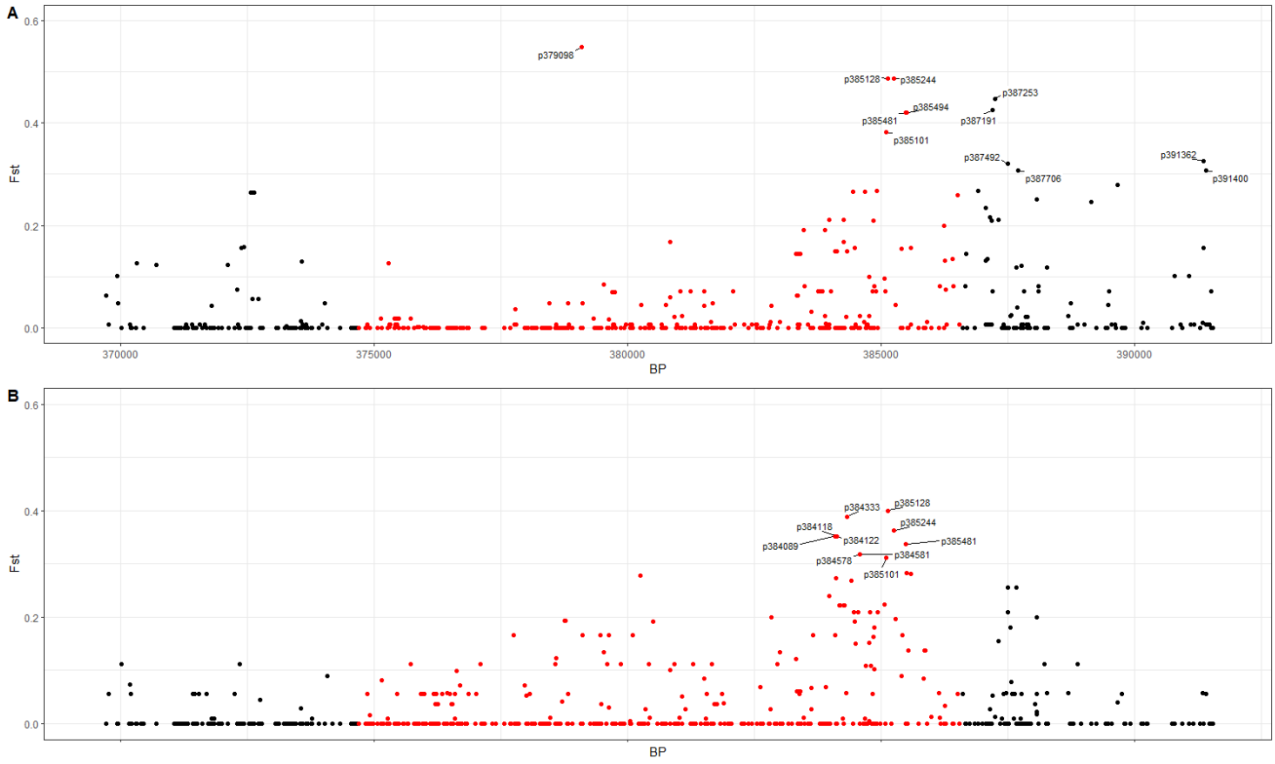

Figura 28. Distribuzione dell' F_{ST} nella regione FSHR ± 5 Kb del contig 2337 nelle popolazioni dell'Egeo (A) e del Tirreno (B). In rosso sono evidenziate le varianti localizzate nel gene fshr.

Allo scopo di individuare le mutazioni distintive tra maschi conformi e non-conformi in comune tra le popolazioni sono stati selezionati SNP presenti in entrambe le popolazioni che avessero F_{ST} >0,1. In questo modo sono stati identificati 16 SNP, tra questi anche le varianti già riscontrate nel precedente lavoro di Ferraresso et al., ora presenti con una denominazione differente in quanto è stato utilizzato un diverso genoma di riferimento (MuCe179, MuCe206 e MuCe322 corrispondono alle varianti 385101, 385128, 385244).

Le 13 nuove varianti identificate (383318, 384089, 384106, 384118, 384122, 384333, 384485, 385481, 385494, 385584, 387313, 387492, 387667) sono risultate essere localizzate in diverse regioni del gene *fshr* (i.e. introni, esoni, 3'UTR) nonché nella regione intergenica a valle del gene, suddivisi secondo la Tabella 7.

In Tabella 7 vengono riportate le frequenze alleliche dell'allele alternativo al genoma di riferimento nelle differenti posizioni identificate e i dati dell'F_{ST}, suddivisi secondo le popolazioni.

Ci si è quindi concentrati, in ciascuna popolazione, sui 20 SNP con F_{ST} più elevato. Facendo l'intersezione tra Egeo e Tirreno, sono risultate essere 6 le varianti in comune tra le due popolazioni. Oltre alle 3 già note dal lavoro precedente di Ferraresso et al. (2021), le rimanenti (posizioni 385481, 385494, 387492) hanno dimostrato essere molto interessanti per la differenziazione tra maschi conformi e non-conformi, con frequenze

alleliche significativamente differenti per tutte e tre le varianti tra individui non-conformi e conformi (Fisher's exact test $p < 0,05$).

Posizione	Tratto genico	Tirreno			Egeo		
		F_{ST}	Freq. Conformi	Freq. Non-conformi	F_{ST}	Freq. Conformi	Freq. Non-conformi
383318	Introne FSHR	0,12	0,75	0,45	0,14	0,95	0,72
384089	Introne FSHR	0,35	0,55	0,10	0,15	0,30	0,06
384106	Introne FSHR	0,27	0,55	0,15	0,15	0,30	0,06
384118	Introne FSHR	0,35	0,55	0,10	0,15	0,30	0,06
384122	Introne FSHR	0,35	0,55	0,10	0,15	0,30	0,06
384333	Introne FSHR	0,39	0,50	0,05	0,15	0,30	0,06
384485	Introne FSHR	0,18	0,60	0,25	0,16	0,40	0,11
385101	**Esone FSHR**	0,31	0,50	0,00	0,38	0,50	0,06
385128	**Esone FSHR**	0,4	0,50	0,00	0,49	0,50	0,00
385244	**Esone FSHR**	0,36	0,55	0,00	0,49	0,50	0,00
385481	Esone FSHR	0,34	0,60	0,10	0,42	0,55	0,06
385494	3'UTR	0,28	0,55	0,10	0,42	0,55	0,06
385584	3'UTR	0,28	0,60	0,95	0,16	0,80	1,00
387313	Intergenico	0,16	0,50	0,80	0,21	0,75	1,00
387492	Intergenico	0,26	0,35	0,75	0,32	0,65	1,00
387667	Intergenico	0,26	0,65	0,25	0,12	0,35	0,11

Tabella 7. Frequenze alleliche nelle differenti posizioni identificate. Gli SNP alla posizione 385101, 385128 e 385244 erano già stati identificati in un precedente lavoro (Ferraresso, et al., 2021). Gli SNP con lo sfondo verde corrispondono ad introni dell'FSHR, gli SNP con lo sfondo giallo corrispondono ad esoni dell'FSHR, gli SNP con lo sfondo blu corrispondono alla regione del 3' UTR, mentre gli SNP con sfondo rosso sono intergenici.

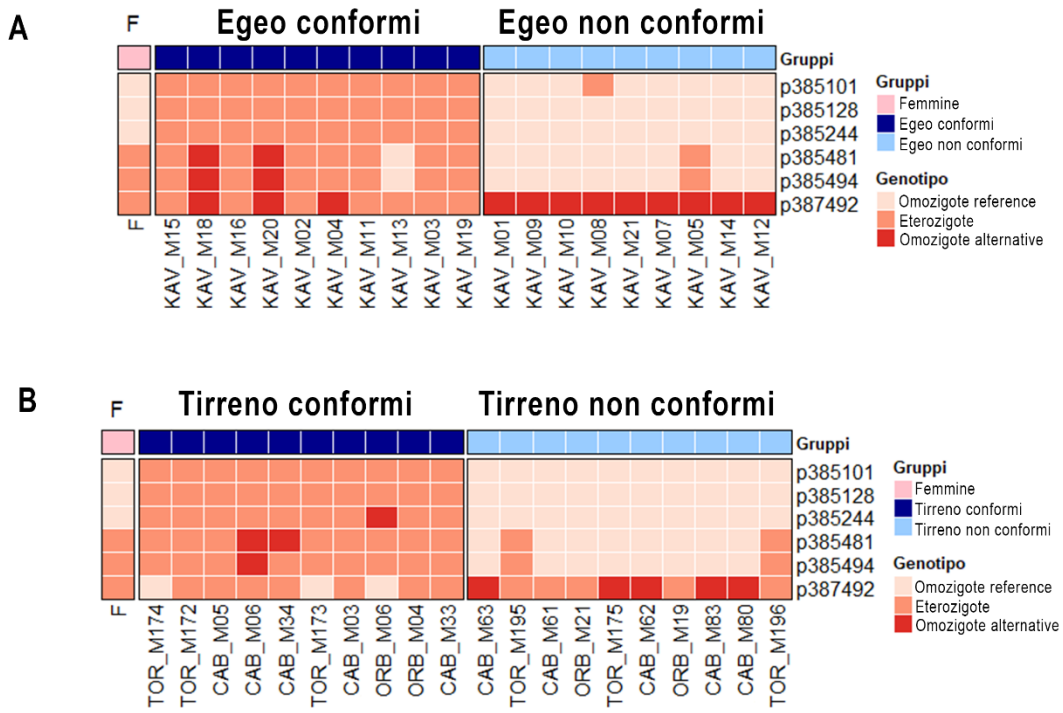

Figura 29. Heatmap dei differenti genotipi nelle 6 posizioni in comune tra i 20 più alti F_{ST} identificati nelle due popolazioni. Il grafico A corrisponde alla popolazione dell'Egeo, il grafico B corrisponde alla popolazione del Tirreno. Lungo le colonne sono presenti i differenti campioni analizzati, la suddivisione tra conformi, non-conformi e genotipo femminile è visualizzata graficamente grazie al codice colore della prima riga. Il genotipo femminile, presente nella prima colonna, deriva dai dati del Pool-Seq del precedente lavoro di Ferraresso et al. (2021).

Nella Figura 29 è possibile visualizzare graficamente i genotipi delle 6 posizioni identificate dall'intersezione dei 20 F_{ST} più elevati delle popolazioni di Egeo e del Tirreno. Per poter avere un confronto con il genotipo femminile si sono utilizzati i dati del Pool-Seq delle femmine del Tirreno del precedente lavoro di Ferraresso et al. (2021). È possibile notare l'evidente differenza tra maschi conformi (eterozigoti) e femmine (omozigoti *reference*) nelle posizioni 385101, 385128 e 385244.

Nella posizione 385481, localizzata nell'ultimo esone, la mutazione presenta una sostituzione di una C con una A. Questa è una mutazione missenso in quanto porta alla variazione dell'aminoacido codificato da serina (TCC) a tirosina (TAC). Lo SNP nelle femmine è presente in eterozigosi, così come nei maschi conformi, mentre nella maggior parte dei casi (8 su 10) la mutazione è assente nei maschi non-conformi.

Nella posizione 385494, localizzata nella regione 3'UTR del gene *fshr*, è presente una marcata eterozigosi nel caso dei maschi conformi, mentre, nella maggior parte dei maschi non-conformi (8 su 10) la mutazione è portata in omozigosi. I due maschi non-conformi che sono eterozigoti per questa mutazione sono gli stessi individui (TOR_M195 e TOR_M196) portatori in eterozigosi della mutazione 385481, mutazione assente nel

resto dei maschi non-conformi (vedi Figura 29). Le femmine in questa posizione sono eterozigoti, come la maggioranza dei maschi conformi.

La mutazione in posizione 387492, localizzata in una regione intergenica a valle dell'*fshr*, è portata in eterozigosi (wt/387492) in 7 maschi conformi su 10, mentre nei maschi non-conformi è presente in omozigosi per il 50% degli individui (l'altra metà è eterozigote). Le femmine in questa posizione sono eterozigoti, come la maggioranza dei maschi conformi.

Grazie all'identificazione delle mutazioni 385481, 385494 e 387492 e alla loro associazione con le mutazioni M1, è possibile differenziare geneticamente maschi conformi e non-conformi, con l'unica eccezione di 2 individui non-conformi del Tirreno, i quali presentano genotipo indistinguibile da quello femminile per questi SNP pur essendo fenotipicamente maschi.

5. Discussione

M. cephalus è una specie gonocorica con determinazione genetica del sesso, tuttavia non presenta cromosomi sessuali eterocromatici e, ad oggi, non è stato definitivamente indicato in bibliografia alcun MSD (Rossi, et al., 1996). Il dimorfismo sessuale non è evidente e la distinzione macroscopica nei due sessi prima della completa maturità sessuale è molto complessa. L'interesse economico legato al cefalo, tuttavia, è connesso principalmente agli individui di sesso femminile le cui gonadi vengono pressate, salate ed essiccate e vendute come "Bottarga", una prelibatezza culinaria che può raggiungere il costo di 300 €/Kg. L'aumentata richiesta globale di questo prodotto ha determinato un forte incremento del valore commerciale di questa specie, tanto da guadagnarsi l'appellativo di "oro grigio" dagli addetti ai lavori (Cossu, et al., 2021). Proprio perché il valore economico di un individuo di sesso femminile è di gran lunga superiore rispetto ad uno di sesso maschile, 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.

Negli ultimi anni, sono stati fatti notevoli passi avanti nell'identificazione di loci legati alla determinazione del sesso in *M. cephalus*. Dor e colleghi (2016) hanno per primi identificato una regione sex-linked di circa 2 Mb nel genoma di *M. cephalus*, hanno inoltre dimostrato che il sesso della progenie risultava essere determinato esclusivamente dagli alleli trasmessi dal padre, con conseguente determinazione del sesso di tipo XY.

Lo stesso gruppo di ricerca (Dor, et al., 2020) ha successivamente indicato i geni *bccip, dhx32a, dock1* e *fshr (gth-ri)* come possibili candidati MSD. Il gene *fshr* è stato identificato anche dagli studi di Ferraresso et al. (2021) e Curzon et al. (2021) come probabile gene legato alla determinazione del sesso in *M. cephalus*, grazie all'identificazione di due mutazioni missenso ed una mutazione sinonima (Ferraresso, et al., 2021) presenti in individui di sesso maschile con frequenza simile a quella che si riscontra nelle varianti sesso-specifiche di tipo XX/XY.

Anche in *Solea senegalensis* è stata recentemente identificata una correlazione tra *fshr* e il sesso fenotipico, con 41 varianti alleliche concordi con il sistema di determinazione sessuale di tipo XX/XY (Herrán, et al., 2022).

Il gene *fshr* codifica per i recettori dell'ormone follicolo-stimolante. L'ormone follicolo-stimolante (FSH) viene secreto dall'ipofisi e, nei vertebrati, si lega ai recettori localizzati sulle cellule della granulosa nell'ovaio e sulle cellule del Sertoli nel testicolo per regolare lo sviluppo delle gonadi e promuoverne la crescita. In medaka (Murozumi, et al., 2014)

e in zebrafish (Zhang, et al., 2015) femmine *knock-out* per *fshr* hanno mostrato un completo fallimento dell'attivazione follicolare e hanno subito un *sex reversal* verso il sesso maschile, mostrando una normale fertilità. I maschi in assenza di *fshr* hanno invece mostrato una normale fertilità, suggerendo un importante ruolo dell'FSH nello sviluppo delle gonadi femminili dei teleostei.

In *M. cephalus* i polimorfismi descritti da Ferraresso et al. (2021) e da Curzon et al. (2021) hanno però mostrato una penetranza incompleta, con una percentuale di maschi "non-conformi" - cioè maschi che mostravano il genotipo wt/wt tipico del sesso femminile - variabile tra popolazioni di origine geografica diversa (8-45%) avanzando quindi l'ipotesi della presenza di altri loci legati alla determinazione del sesso.

Il presente lavoro di tesi si è quindi incentrato nella ricerca di ulteriori variabili genetiche in grado di differenziare i maschi dalle femmine di *M. cephalus.* Il sequenziamento a medio coverage dell'intero genoma di maschi conformi e maschi non-conformi di 2 popolazioni, Egeo e Tirreno, ha permesso di superare alcune limitazioni degli studi precedenti. Il sequenziamento di singoli individui ha permesso di assegnare in modo preciso il genotipo di ciascun animale, obiettivo che con Pool-seq non era stato possibile raggiungere. Il sequenziamento dell'intero genoma offre poi la possibilità di identificare polimorfismi in qualunque regione del genoma, e quindi anche in loci genetici diversi dalla regione candidata presa in esame.

In questo lavoro, l'analisi della regione FSHR±5kb ha permesso di identificare 3 nuove varianti, 2 nel gene fshr e una nell'immediata regione a valle, che presentano una frequenza allelica significativamente diversa tra maschi conformi e maschi non-conformi e, in aggiunta, permette di differenziare i maschi non-conformi dalle femmine.

Gli SNP in posizione 385481 e 385494, rispettivamente localizzati nell'ultimo esone e nella regione 3' UTR del gene fshr risultano essere presenti (in eterozigosi o, addirittura, in omozigosi) nel 95% (19/20) dei maschi conformi analizzati mentre solo il 16% (3/19) dei maschi non-conformi presenta l'allele mutato.

Lo SNP in posizione 387492 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, il che permetterebbe di differenziarli in modo certo dalle femmine che presentano invece la mutazione in eterozigosi. Il quadro è invece diverso nei maschi non-conformi del Tirreno, nei quali la mutazione è presente in omozigosi nel 50% dei casi mentre 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.

Come già riportato, prendendo in considerazione i soli 3 polimorfismi identificati nello studio di Ferraresso et al. (2021) una percentuale di individui, che arriva a sfiorare il 50% nell'Egeo, anche se fenotipicamente maschi, verrebbe erroneamente classificata come femmine.

Mediante i dati raccolti nel lavoro di questa tesi è possibile affermare che con l'utilizzo dei 2 nuovi SNP identificati in posizione 385481 e 385494 è possibile classificare correttamente l'80% (8/10) degli individui non-conformi del Tirreno e l'89% (8/9) degli individui non-conformi dell'Egeo. La messa a punto di un test genetico basato sull'amplificazione ed il sequenziamento di una a sola regione di 400 bp, contenente i 5 SNP 385101, 385128, 385244, 385481 e 385494, permetterebbe quindi di aumentare l'accuratezza nell'identificazione del sesso di *M. cephalus* dal 88% al 98% per il Tirreno e dal 55% al 95% per l'Egeo.

Queste 5 variabili hanno un profilo ben definito nelle femmine, le prime 3 posizioni (385101, 385128, 385244) non presentano gli SNP identificati mentre le seconde due (385481 e 385494) presentano gli SNP in eterozigosi.

Riscontrando invece un profilo eterozigote per 385101, 385128 e 385244, oppure omozigote *(wt/wt* + *mut/mut)* in 385481 e 385494 si ha una probabilità del 92% (36/39, nel campione utilizzato in questo lavoro di tesi) che l'individuo in questione sia maschio.

L'accuratezza aumenterebbe ulteriormente, per la popolazione dell'Egeo, includendo anche lo SNP in posizione 387492. La posizione 387492 è fisicamente molto vicina agli altri SNP identificati; quindi, pur essendo un SNP identificato come intergenico e a valle dell'*fshr* potrebbe essere in *linkage* con le altre mutazioni. Prendendo in considerazione anche quest'ultimo polimorfismo, creando così un test genetico che utilizzi 6 SNP, si riuscirebbe a identificare geneticamente come maschi la totalità degli individui dell'Egeo analizzati nel presente lavoro. Questa ipotesi va però verificata poiché il sequenziamento delle femmine dell'Egeo si è limitato alle prime 3 varianti (385101, 385128, 385244) in base alle quali il genotipo delle femmine del Tirreno e quello delle femmine dell'Egeo è risultato essere concorde. Alla luce però della variabilità di popolazione osservata anche nel presente lavoro non è al momento possibile affermare con certezza che questo valga anche per gli SNP in posizione 385481, 385494 e 387492. Il sequenziamento delle femmine e di ulteriori maschi (conformi e non-conformi) è quindi necessario per poter corroborare i risultati ottenuti in questo lavoro di tesi.

La presenza, tuttavia, di una bassa percentuale di maschi (2-5%) che ancora non verrebbe identificata con le 5 varianti soprariportate e la differenza osservata tra le popolazioni di Egeo e Tirreno suggeriscono, inoltre, che possa contribuire anche una

componente geografica nella determinazione del sesso in *M. cephalus.* Una possibile ipotesi potrebbe essere un comportamento simile a quello identificato nella spigola (*Dicentrarchus labrax*). *D. labrax* è una specie in cui è presente sia ESD che GSD e recentemente sono stati identificati una serie di loci con un effetto minore ma che contribuiscono ugualmente alla determinazione sessuale (PSD – *polygenic sex determination*) in questa specie. Tuttavia, questi loci sono parzialmente differenti in base all'origine geografica dei soggetti presi in esame (Faggion, et al., 2019) e non è possibile escludere che questo non sia possibile anche per *M. cephalus*.

È noto che anche in altre specie la determinazione del sesso sia poligenica (Anderson, et al., 2012) e che possa esserci l'intervento anche di una componente ambientale. Per questo motivo è necessaria l'estensione dell'analisi all'intero genoma, e non solo alla regione *fshr*, di un maggiore numero di individui di sesso maschile di differenti popolazioni, così da poter verificare l'esistenza di eventuali loci con un minor effetto ma che integrano l'azione del locus predominante con eventuali differenze a seconda dell'origine geografica.

È possibile ipotizzare che nella regione terminale del gene *fshr* in *M. cephalus* gli SNP identificati alle posizioni 385101, 385128 e 385244 costituiscano la mutazione principale che ne determina il sesso svolgendo un ruolo come dominante negativa, causando cioè la perdita della funzione della copia rimanente del gene e inducendo lo sviluppo maschile, così come accade in zebrafish (Zhang, et al., 2015) e medaka (Murozumi, et al., 2014) con *fshr* KO.

Nei maschi non-conformi invece ciò non si verifica e la presenza degli SNP 385481 e 385494 andrebbe a confutare questa ipotesi, essendo presenti in omozigosi nei maschi non-conformi ed in eterozigosi nelle femmine e rendendo evidente ancora una volta la necessità dell'estensione dell'analisi all'intero genoma di individui di sesso maschile e femminile provenienti da differenti aree geografiche di *M. cephalus.*

6. Conclusione

Il presente lavoro di tesi ha permesso di identificare nuove mutazioni a carico del gene *fshr* e nell'immediata regione a valle, in grado di aumentare notevolmente l'accuratezza nella identificazione del sesso di *M. cephalus*.

Le mutazioni alle posizioni 385481 e 385494 nel gene *fshr* se associate alle mutazioni M1 (SNP 385101, 385128 e 385244), possono permettere la creazione di un test genetico basato sull'amplificazione ed il sequenziamento di una sola regione di 400 bp, contenente i 5 SNP capaci di differenziare geneticamente con alta probabilità maschi da femmine. In questo modo sarà possibile sessare precocemente - e con maggiore accuratezza rispetto all'utilizzo di soli 3 SNP M1 - gli individui di *M. cephalus.*

In aggiunta, la mutazione alla posizione 387492 potrebbe permettere, in combinazione con le altre 5 sopra descritte (385101, 385128, 385244, 385481, 385494), di aumentare ulteriormente l'accuratezza dell'identificazione genetica dei maschi, almeno nella popolazione dell'Egeo.

In questo modo sarebbe possibile, attraverso un singolo test dal costo di circa 5€, identificare precocemente il sesso degli individui, così da poter portare in accrescimento una popolazione in grande maggioranza (se non esclusivamente) femminile. Questo potrebbe dimostrarsi un importante passo avanti nel migliorare/ottimizzare il rendimento economico dell'allevamento dei cefali mirati alla produzione di bottarga il cui prezzo può raggiungere sul mercato dai 150 ai 300 €/kg mentre per un cefalo adulto di taglia commerciale il prezzo è di gran lunga inferiore (intorno ai 4 €/Kg).

Le mutazioni riscontrate nel locus *fshr* hanno dato informazioni importanti sul controllo genetico della determinazione del sesso in *M. cephalus* e fornito strumenti molto utili per lo sviluppo di test molecolari per il sessaggio precoce. Tuttavia, i risultati ottenuti hanno rivelato un quadro assai complesso. Al momento, non è stato infatti possibile identificare un aplotipo univoco che differenzi al 100% il sesso fenotipico maschile da quello femminile lasciando aperta la possibilità dell'esistenza di altri loci e/o di una componente ambientale che influenzino la determinazione del sesso in *M. cephalus*. Questo rende necessario estendere l'analisi all'intero genoma nonché ad un numero maggiore di esemplari includendo un numero adeguato di femmine sia del Tirreno che dell'Egeo e, possibilmente, includendo nelle analisi anche individui provenienti da altre popolazioni di diversa origine geografica.

7. Ringraziamenti

Vorrei ringraziare innanzitutto il Prof. Tomaso Patarnello del Dipartimento di Biomedicina Comparata e Alimentazione (BCA), che mi ha saputo indirizzare allo svolgimento del mio lavoro di tesi presso i laboratori del dipartimento di BCA e della sua continua disponibilità. Ringrazio anche la Dott.ssa Serena Ferraresso, che mi ha seguita e guidata durante tutta la mia attività con competenza e professionalità. Ringrazio il Prof. Luca Bargelloni per gli spunti di riflessione in questo intero progetto di tesi. Ringrazio il Dott. Massimiliano Babbucci e il Dott. Mbarsid Racaku per il loro aiuto nelle analisi informatiche. Ringrazio infine la Dott.ssa Rafaella Franch e la Dott.ssa Giulia Dalla Rovere per la loro disponibilità nelle attività di laboratorio.

8. Appendice

Campione	*Raw reads*	*Trimmed*	% mantenuta
1-KAV-M17	73.705.219	71.993.158	97,68%
2-CAB-M61	56.728.135	54.684.008	96,40%
3-KAV-M08	81.691.894	79.325.303	97,10%
4-KAV-M19	86.777.622	84.477.496	97,35%
5-CAB-M83	63.036.929	61.183.793	97,06%
6-CAB-M33	58.659.499	57.186.247	97,49%
7-CAB-M80	61.304.304	59.756.285	97,47%
8-KAV-M12	79.661.092	77.643.562	97,47%
9-TOR-M196	104.570.944	102.255.355	97,79%
10-TOR-M174	62.572.060	61.089.893	97,63%
11-CAB-M63	78.425.850	76.503.611	97,55%
12-KAV-M01	103.412.483	100.921.700	97,59%
13-TOR-M172	77.801.275	76.156.209	97,89%
14-KAV-M15	94.393.578	92.204.287	97,68%
15-CAB-M05	90.768.899	88.738.209	97,76%
16-CAB-M06	87.249.332	85.268.651	97,73%
17-TOR-M195	64.686.547	63.412.237	98,03%
18-KAV-M09	71.207.629	69.748.777	97,95%
19-CAB-M34	69.192.400	67.813.601	98,01%
20-TOR-M173	64.688.303	63.454.471	98,09%
21-CAB-M03	64.405.962	63.224.439	98,17%
22-KAV-M18	76.864.396	75.412.121	98,11%
23-ORB-M06	76.304.636	74.859.083	98,11%
24-ORB-M21	62.078.332	61.017.853	98,29%
25-TOR-M175	80.855.017	78.500.892	97,09%
26-ORB-M04	61.391.709	60.229.009	98,11%
27-KAV-M16	66.087.577	64.861.847	98,15%
28-KAV-M10	86.140.070	84.311.940	97,88%
29-KAV-M20	72.177.739	70.688.630	97,94%

30-KAV-M02	86.736.781	84.743.137	97,70%
31-CAB-M62	67.734.022	66.075.912	97,55%
32-KAV-M21	67.560.374	66.354.460	98,22%
33-KAV-M04	67.914.082	66.346.871	97,69%
34-ORB-M19	61.667.591	60.313.097	97,80%
35-KAV-M07	72.629.819	71.082.672	97,87%
36-KAV-M11	70.071.651	68.554.311	97,83%
37-KAV-M13	59.741.433	58.303.233	97,59%
38-KAV-M05	75.069.908	73.517.701	97,93%
39-KAV-M03	68.121.062	66.662.065	97,86%
40-KAV-M14	71.681.040	70.099.939	97,79%

Appendice A. Confronto tra raw reads e reads mantenute dopo il trimming

9. Bibliografia

Anderson, J. et al., 2012. Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. *PLoS One,* 7(7), e40701.

Arkhipchuk, V., 1995. Role of chromosomal and genome mutations in the evolution of bony fishes. *Hydrobiologia Journal,* 55-65.

Bekhit, A. E.-D., 2022. *Fish Roe: Biochemistry, Products, and Safety.* 1st edition ed. :Academic Press.

Binelli, G. e Ghisotti, D., 2017. *Genetica.* I/2017 ed. :Edises.

Capel, B., 2017. Vertebrate sex determination: evolutionary plasticity of a fundamental switch. *Nat Rev Genet,* 18, 675-689.

Chadwick, D. e Goode, J., 2002. *The Genetics and Biology of Sex Determination.* 1° ed. :Wiley.

Chen, J., Zhu, Z. e Hu, W., 2022. Progress in research on fish sex determining genes. *Water Biology and Security,* 1(1), 100008.

Chen, S., Zhang, G., Shao, C. et al., 2014. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nat Genet ,* 46, 253-260.

Cortez, D. et al., 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature,* 508, 488-493.

Cossu, P. et al., 2021. Genetic patterns in Mugil cephalus and implications for fisheries and aquaculture management. *Sci Rep,* 11(1), 2887.

Crosetti, D., 2016. *Current state of grey mullet fisheries and culture. In Biology, Ecology and Culture of Grey Mullet (Mugilidae)*:CRC Press.

Curzon, A. et al., 2021. A novel c.1759T>G variant in follicle-stimulating hormone-receptor gene is concordant with male determination in the flathead grey mullet (Mugil cephalus). *G3 (Bethesda),* 11(2), kaa044.

Diciotti, R. et al., 2022. Use of otoliths for estimating age of Mugil cephalus L. destined to “bottarga” production in Tortolì lagoon (central western Sardinia, western Mediterranean).

Dor, L. et al., 2020. Preferential Mapping of Sex-Biased Differentially-Expressed Genes of Larvae to the Sex-Determining Region of Flathead Grey Mullet (Mugil cephalus). *Frontiers in Genetics,* 11, 839.

Dor, L. et al., 2016. Identification of the sex-determining region in flathead grey mullet (Mugil cephalus). *Anim Genet,* 47(6), 698-707.

Eggert, C., 2004. Sex determination: the amphibian models. *Reproduction Nutrition Development,* 44(6), 539-549.

Elshire, R. et al., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One,* 6(5), e19379.

Eppley, S. M. e Jesson, L. K., 2008. Moving to mate: the evolution of separate and combined sexes in multicellular organisms. *Journal of Evolutionary Biology,* 21, 727-736.

Faggion, S. et al., 2019. Population-specific variations of the genetic architecture of sex determination in wild European sea bass Dicentrarchus labrax L.. *Heredity,* 122(5), 612-621.

Ferraresso, S. et al., 2021. fshr: a fish sex-determining locus shows variable incomplete penetrance across flathead grey mullet populations. *iScience,* 24(1), 101886.

Gamble, T. et al., 2015. Restriction Site-Associated DNA Sequencing (RAD-seq) Reveals an Extraordinary Number of Transitions among Gecko Sex-Determining Systems. *Molecular Biology and Evolution,* 32(5), 1296-1309.

García-López, et al., 2010. Studies in zebrafish reveal unusual cellular expression patterns of gonadotropin receptor messenger ribonucleic acids in the testis and unexpected functional differentiation of the gonadotropins. *Endocrinology,* 151(5), 2349-2360.

Ge, C. et al., 2017. Dmrt1 induces the male pathway in a turtle species with temperature-dependent sex determination. *Development,* 144(12), 2222-2233.

Hattori, R. et al., 2012. A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. *Proc Natl Acad Sci USA,* 109(8), 2955-2959.

Herpin, A. e Schartl, M., 2015. Plasticity of gene-regulatory networks controlling sex determination: of masters, slaves, usual suspects, newcomers, and usurpators. *EMBO reports,* 16(10), 1260-1274.

Herrán, R. D. l. et al., 2022. A chromosome-level genome assembly enables the identification of the follicle stimulating hormone receptor as the master sex determining gene in Solea senegalensis. *bioRxiv, 2022.03.02,*482245.

Holsinger, K. e Weir, B., 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat Rev Genet,* 10, 639-650.

Hu, T., Chitnis, N., Monos, D. e Dinh, A., 2021. Next-generation sequencing technologies: An overview. *Human Immunology,* 82(11), 801-811.

Jombart, T., Devillard, S. e Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet,* 11(94), 1471-2156.

Kamiya, T., Kai, W., Tasumi, S. et al., 2012. A Trans-Species Missense SNP in Amhr2 Is Associated with Sex Determination in the Tiger Pufferfish, Takifugu rubripes (Fugu). *PLOS Genetics,* 8(7), e1002798.

Kondo, M. et al., 2003. Absence of the candidate male sex-determining gene dmrt1b(Y) of medaka from other fish species. *Curr Biol.,* 13(5), 416-420.

Koopman, P. et al., 1991. Male development of chromosomally female mice transgenic for Sry. *Nature,* 351, 117-121.

Lambeth, L. S. et al., 2014. Over-expression of DMRT1 induces the male pathway in embryonic chicken gonads. *Developmental Biology,* 389(2), 160-172.

Li, M. et al., 2015. A Tandem Duplicate of Anti-Müllerian Hormone with a Missense SNP on the Y Chromosome Is Essential for Male Sex Determination in Nile Tilapia, Oreochromis niloticus. *PLOS Genetics ,* 11(11), e1005678.

Li, X. Y. e Gui, J. F., 2018. Diverse and variable sex determination mechanisms in vertebrates. *Science China. Life sciences,* 61(12), 1503-1514.

Lou, R. N., Jacobs, A., Wilder, A. P. e Therkildsen, N. O., 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology,* 30(23), 5966-5993.

Matsuda, M., 2005. Sex Determination in the Teleost Medaka, Oryzias latipes. *Annu Rev Genet,* 39, 293-307.

Matsuda, M. et al., 2002. DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature,* 417(6888), 559-563.

Matsuda, M. et al., 2003. Oryzias curvinotus has DMY, a gene that is required for male development in the medaka, O. latipes. *Zoolog Sci.,* 20(2), 159-161.

McDonough, C., Roumillat, W. e Wenner, C., 2005. Sexual differentiation and gonad development in striped mullet (Mugil cephalus L.) from South Carolina estuaries. *Fishery Bulletin,* 103, 601-619.

Miura, I., 2017. Sex Determination and Sex Chromosomes in Amphibia. *Sexual development : genetics, molecular biology, evolution, endocrinology, embryology, and pathology of sex determination and differentiation,* 11(5-6), 298-306.

Moore, E. C. e Roberts, R. B., 2013. Polygenic sex determination. *Current Biology,* 23(12), R510-R512.

Munday, P., Buston, P. e Warner, R., 2006. Diversity and flexibility of sex-change strategies in animals. *Trends in Ecology & Evolution,* 21(2), 89-95.

Murdock, C. e Wibbels, T., 2006. Dmrt1 expression in response to estrogen treatment in a reptile with temperature-dependent sex determination. *Journal of experimental zoology. Part B, Molecular and developmental evolution,* 306(2), 134-139.

Murozumi, N. et al., 2014. Loss of Follicle-Stimulating Hormone Receptor Function Causes Masculinization and Suppression of Ovarian Development in Genetically Female Medaka. *Endocrinology,* 155(8), 3136-3145.

Myosho, T. et al., 2012. Tracing the emergence of a novel sex-determining gene in medaka, Oryzias luzonensis. *Genetics,* 191(1), 163-170.

Nef, S. e Vassalli, J.-D., 2009. Complementary pathways in mammalian female sex determination. *Journal of Biology ,* 8(8), 74.

Nelson, J., Grande, T. e Wilson, M., 2016. *Fishes of the World, Fifth Edition* :Wiley.

Olukolajo, S. O., 2017. Reproductive studies of striped mullet (Mugil cephalus) from high brackish tropical lagoon. *Livestock Research for Rural Development,* 29(156).

Pennell, M. W., Mank, J. E. e Peichel, C. L., 2018. Transitions in sex determination and sex chromosomes across vertebrate species. *Molecular ecology,* 27(19), 3950-3963.

Raymond, J. A. J. et al., 2022. Comparative genome size estimation of different life stages of grey mullet, Mugil cephalus Linnaeus, 1758 by flow cytometry. *Aquaculture Research,* 53, 1151-1158.

Reichwald, K. et al., 2015. Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell,* 163(6), 1527-1538.

Rens, W. et al., 2004. Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution. *Proc Natl Acad Sci USA,* 101(46), 16257-16261.

Rens, W. et al., 2007. The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biol,* 8, R243.

Rossi, A. r., Crosetti, D., Gornung, E. e Luciana Sola, 1996. Cytogenetic analysis of global population of Mugil cephalus (striped mullet) by different staining techniques and fluorescent in situ hybridization. *Heredity,* 76, 77-82.

Sadovy, Y. e Shapiro, D. Y., 1987. Criteria for the diagnosis of hermaphroditism in fishes. *Copeia,* 1987(1), 136-156.

Sánchez, L. e Chaouiya, C., 2018. Logical modelling uncovers developmental constraints for primary sex determination of chicken gonads. *J R Soc Interface,* 15(142), 20180165.

Sanger, F., Nicklen, S. e Coulson, A., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74(12), 5463-5467.

Shaw, C.-M. e Daigee, H., 2006. The Impact of Upstream Catch and Global Warming on the Grey Mullet Fishery in Taiwan: A Non-cooperative Game Analysis. *Marine Resource Economics,* 21(3), 285-300.

Sinclair, A. H. et al., 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature,* 346, 240-244.

Smith, C. A. et al., 1999. Conservation of a sex-determining gene. *Nature,* 402, 601-602.

Smith, C. et al., 2009. The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature,* 461, 267-271.

Sutou, S., Mitsui, Y. e Tsuchiya, K., 2001. Sex determination without the Y Chromosome in two Japanese rodents Tokudaia osimensis osimensis and Tokudaia osimensis s *Mammalian Genome,* 12(1), 17-21.

Takehana, Y. et al., 2014. Co-option of Sox3 as the male-determining factor on the Y chromosome in the fish Oryzias dancena. *Nature Communications,* 5, 4157.

Thépot, D., 2021. Sex Chromosomes and Master Sex-Determining Genes in Turtles and Other Reptiles. *Genes,* 12(11), 1822.

Thomson, J., 1963. *Fisheries and Oceanography; Grey mullet; Fisheries synopsis.* Cronulla, Sydney: Division of Fisheries and Oceanography. CSIRO.

Wallis, M. C. et al., 2007. Sex determination in platypus and echidna: autosomal location of SOX3 confirms the absence of SRY from monotremes. *Chromosome Res ,* 15, 949.

Whitfield, A. K., Panfili, J. e Durand, J.-D., 2012. A global review of the cosmopolitan flathead mullet Mugil cephalus Linnaeus 1758 (Teleostei: Mugilidae),with emphasis on the biology, genetics, ecology and fisheries aspects of this apparent species complex. *Rev Fish Biol Fisheries ,* 22, 641-681.

Yano, A. et al., 2012. An Immune-Related Gene Evolved into the Master Sex-Determining Gene in Rainbow Trout, Oncorhynchus mykiss. *Curr Biol,* 22(15), 1423-1428.

Yoshimoto, S. et al., 2008. A W-linked DM-domain gene, DM-W, participates in primary ovary development in Xenopus laevis. *Proc Natl Acad Sci USA,* 105(7), 2469-2474.

Zhang, F. e Lupski, J. R., 2015. Non-coding genetic variants in human disease. *Human molecular genetics,* 24(R1), R102-10.

Zhang, Z., Lau, S.-W., Zhang, L. e Ge, W., 2015. 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,* 156(10), 3747-3762.

10. Sitografia

http://www.bio-bwa.sourceforge.net/

https://www.fao.org/

https://www.fishbase.se

https://www.illumina.com/

http://www.ittiofauna.org/