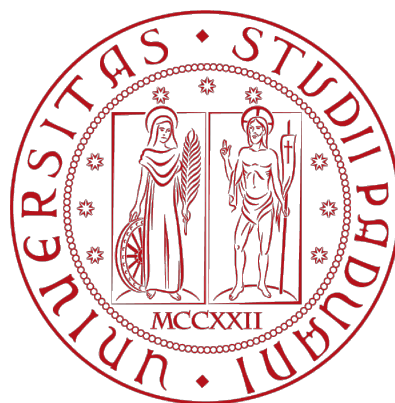




UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Biomedicina Comparata e Alimentazione

Dipartimento di Agronomia, Animali, Alimenti, Risorse naturali e
Ambiente

Corso di laurea magistrale in
Biotecnologie per l'alimentazione

Identificazione nell'intero genoma delle variazioni del numero di copie (CNV) nella razza caprina Girgentana e validazione con tecnologie di sequenziamento di ultima generazione

Relatore

Prof. Martino Cassandro

Correlatore

Dott. Baldassare Portolano

Laureando
Antonio Miceli

Matricola n.
1147989

ANNO ACCADEMICO 2017/2018

Indice

RIASSUNTO	1
ABSTRACT	2
Capitolo I: Introduzione	3
1. La capra Girgentana.....	4
1.1 Cenni storici.....	4
1.2 Caratteristiche fenotipiche, produttive e riproduttive	5
1.3 Salvaguardia e valorizzazione della biodiversità zootecnica.....	7
2 I marcatori molecolari	9
2.1 Descrizione generale e caratteristiche	9
2.2 <i>Single Nucleotide Polymorphisms</i> o SNPs.....	10
2.3 Principali utilizzi dei marcatori SNPs nel settore zootecnico	12
3 Le Copy Number Variants (CNVs).....	13
3.1 Definizione ed identificazione tramite diverse metodologie.....	13
3.2 Importanza delle CNVs nelle specie di interesse zootecnico	15
3.2.1 CNVs nei bovini	15
3.2.2 CNVs negli ovini.....	17
3.2.3 CNVs nei caprini	17
3.3 Validazione delle CNVs con differenti metodi.....	18
Capitolo II: Parte Sperimentale.....	20
1 Obiettivo della tesi.....	21
2 Materiale e metodi	22
2.1 Genotipizzazione di capre di razza Girgentana con BeadChip Illumina ed analisi bioinformatiche tramite software PennCNV e SVS	22
2.1.1 Campionamento e preparazione del campione	22
2.1.2 I software GenomeStudio 2.0, PennCNV e SVS 8.7.0	23
2.2. Preparazione delle librerie genomiche, sequenziamento ed analisi bioinformatiche per la validazione delle CNVs.....	26
2.2.1 Preparazione della libreria per il sequenziamento	26
2.2.2 Bioinformatica in NGS.....	27

2.2.2.1 Formato FASTQ.....	28
2.2.2.2 Controllo Qualità	29
2.2.2.3 Trimming ed allineamento delle reads al genoma di riferimento	30
2.2.2.4 CONTROL-FREEC	32
2.2.3 Annotazione di geni all'interno delle CNVRs validate.....	33
Capitolo III: Risultati e discussione	34
1 Risultati CNVs provenienti da dati di genotyping	35
1.1 Risultati PennCNV e SVS	35
1.2 CNVRs e definizione di consensus CNVRs	36
2 Dati di sequenziamento	42
2.1 Processamento delle reads e Quality Check	42
2.2 Mappaggio e allineamento	43
2.3 Chiamata e Validazione delle CNVs.....	44
3 Annotazione genica	48
Capitolo IV: Conclusioni	51
Bibliografia.....	53

RIASSUNTO

La razza caprina Girgentana facente parte delle capre domestiche (*Capra hircus* Linnaeus, 1758), trae la sua denominazione da “Girgenti”. Per la specie caprina vi è un numero limitato di studi eseguiti a livello genomico.

L’obiettivo della presente tesi è stato quello di identificare le *copy-number variations* (CNVs) presenti all’interno del genoma della razza Girgentana sulla base dei segnali di intensità (log R ratio) e della frequenza dell’allele B di questi marcatori utilizzando i GoatSNP50 BeadChip Illumina, contenenti 53.347 SNPs. L’identificazione delle CNVs è stata effettuata con l’utilizzo di due differenti software, il PennCNV e l’SVS 8.7.0, ed infine le CNVs sono state aggregate in *CNV regions* (CNVRs). PennCNV ha identificato 192 CNVs mentre SVS ne ha identificate 804. Queste CNVs sono state aggregate in CNVRs. Le CNVs di PennCNV sono state aggregate in 59 CNVRs, mentre le CNVs di SVS sono state aggregate in 361 CNVRs. Le CNVRs comuni ai due software sono state identificate utilizzando la funzione *intersect* di BEDTools che identifica solo le CNVRs presenti nei due datasets e che presentano una sovrapposizione completa (99%). Attraverso questa funzione di BEDTools sono state trovate solo 15 CNVRs *consensus* ovvero comuni tra i due data set. Infine queste sono state validate tramite sequenziamento individuale con tecnologie *high-throughput Next-Generation Sequencing* il quale ha permesso la validazione di 5 CNVRs sul totale delle 15 CNVRs individuate, andando quindi a validare il 33,3% delle CNVRs *consensus*. Ciò rappresenta un risultato notevole rispetto a quanto riscontrato nei lavori precedenti.

Attraverso un’analisi di *gene enrichment* sono stati trovati molti geni che svolgono un ruolo importante nell’adattamento o nella resistenza alle malattie, nei caratteri fenotipici e relativi alla produzione, e nei meccanismi regolatori interni.

Infine, i nostri risultati hanno fornito delle informazioni significative per la costruzione di una mappa più completa delle CNVs del genoma caprino. Questo lavoro, inoltre offre anche una risorsa importante per gli studi futuri che verranno fatti su questa specie per arricchire ancora di più il genoma di questo animale poco studiato.

ABSTRACT

The Girgentana goat breed, which is part of the domestic goats (*Capra hircus* Linnaeus, 1758), derives its name from "Girgenti" (as Agrigento). For the caprine species there is a limited number of studies performed at the genomic level.

The aim of this work was to identify copy number variations (CNVs) present within the genome of the Girgentana goat breed on the basis of the intensity signals (log R ratio) and the frequency of B allele of each marker (BAF), using Illumina GoatSNP50 BeadChip, that containing 53,347 SNPs. The identification of the CNVs was performed with the use of two different software, the PennCNV and the SVS 8.7.0, and after the CNVs were aggregated in the CNV regions (CNVRs). PennCNV identified 192 CNVs while SVS identified 804 CNVs. These CNVs were aggregated in CNVRs. The CNVs of PennCNV were aggregated in 59 CNVRs, while the CNVs of SVS were aggregated in 361 CNVRs. The CNVRs common to both software (*consensus*) were identified with the *intersect* function of BEDTools, that identify only the CNVRs that fully overlap each other. Only 15 *consensus* CNVRs were found between the two data sets. Finally, these were validated through individual sequencing with high-throughput technologies Next-Generation Sequencing which allowed the validation of 5 CNVRs on the total of the 15 identified CNVRs, thus going to validate 33.3% of the CNVRs. This represents a remarkable result and greater than that performed in the previous works.

Through a gene enrichment analysis, many genes have been found that play an important role in adaptation or resistance to disease, in phenotypic traits and related to production and internal regulatory mechanisms. Finally, our results have provided significant information useful for the construction of a more complete map of the CNVs of the goat genome. This work also offers an important resource for future studies that will be done on this species to enrich even more the genome of this less studied animal.

Capitolo I: Introduzione

1. La capra Girgentana

1.1 Cenni storici

La capra domestica (*Capra hircus* Linnaeus, 1758 o, per alcuni autori, *Capra aegagrus hircus*) è il discendente addomesticato dell'egagro (*Capra aegagrus*) dell'Asia Minore e dell'Est europeo.

La razza caprina Girgentana (Figura 1), facente parte delle capre domestiche (*Capra hircus* Linnaeus, 1758), trae la sua denominazione da “Girgenti”, oggi Agrigento, capoluogo della omonima provincia e, dove un tempo, essa segnava la maggiore densità numerica.



Figura 1: capre di razza Girgentana

L'allevamento caprino ha origini antichissime, infatti la capra è tra gli animali domestici più antichi. La sua domesticazione avvenne verso il 9.000-10.000 a.C. nel Medio Oriente a partire dalla *Capra aegagrus* (Bezoar), l'unica vivente all'epoca in quest'area.

Le origini delle capra sono tra le più incerte e ciò causa anche discussioni e controversie. Diversi studi archeologici hanno dimostrato che le capre furono probabilmente prima addomesticate nella regione della Mezzaluna Fertile circa 10.000 anni fa (Zeder M.A. 2000). I primi ad addomesticare le capre, stando ai reperti, furono gli abitanti dei Monti Zagros, in Iran, allo scopo di avere una fonte sicura e sempre accessibile di carne, latte e pelli. Le prime capre domestiche erano tenute in piccoli greggi seminomadi che si spostavano nelle zone collinari alla ricerca di cibo, sempre sorvegliati dai pastori, generalmente giovani o adolescenti (Sardina M.T. 2006).

Mannen e coll. (Mannen H. 2001) hanno suggerito che almeno due specie selvatiche del genere *Capra* hanno contribuito al patrimonio genetico di capre domestiche, mentre un secondo evento di domesticazione potrebbe aver dato origine alle razze di capre tipo Cashmere in Pakistan (Meadow R.H. 1993). A tal proposito, si suppone che l'antenata della capra Girgentana sia la Markhor (*Capra falconeri* Wagner, 1839) o Falconeri (Figura 2) così detta da Falconer, naturalista inglese che per primo la vide nel Kashmir, nell'Afghanistan settentrionale e nel Belucistan.



Figura 2: *Capra falconeri* Wagner, 1839

Intorno al 1960 si asseriva anche che il pelame e le corna della razza Girgentana ricordassero i soggetti asiatici viventi ancora allo stato quasi selvatico, ragion per cui non si dovesse scartare l'ipotesi che la sua origine fosse da ricercarsi tra le capre del Tibet, precisamente nella zona dell'Himalaya.

Amscheler, agronomo tedesco e ricercatore zootecnico, sosteneva che questa capra potesse provenire da una sottospecie della capra Prisca.

Quindi come si può evincere l'origine di questa razza costituisce tutt'ora uno degli interrogativi che gli studiosi si pongono tant'è che essi stessi sono molte volte in disaccordo.

1.2 Caratteristiche fenotipiche, produttive e riproduttive

La capra Girgentana è nota soprattutto per le alte corna a “cavaturacciolo” ovvero a cavatappi, caratteristica fenotipica che la contraddistingue da altre razze; oltre alle corna presenta una barba molto lunga ed estesa a tutta la gola, mantello bianco candido con delle macchie marroncine sul collo e nella testa (Figura 3).



Figura 3: primo piano del viso e delle corna a “caturacciolo”

Questa razza è di taglia media, con la fronte e i mascellari di colore fulvo tendenti al roano e raramente al grigio, spesso caratterizzato da una picchiettatura (soggetti piperini) più o meno estesa. Il pelo è ruvido e di lunghezza media; la pelle bianco-rosea, a volte presenta una leggera pigmentazione. La capra misura al garrese 60-80 cm, mentre il becco (esemplare maschio della *Capra hircus*, detto anche capro, caprone o più raramente irco) arriva a 85 cm, con lunghezza del tronco di 1,06 m. Il peso vivo negli adulti è di 65 Kg nei maschi e 46 Kg nelle femmine; i capretti alla nascita pesano circa 3,5 Kg ed a 60 giorni raggiungono il peso di circa 10 Kg (<https://www.capre.it/le-razze-in-italia/registro-anagrafico-caprino/12-girgentana.html>).

	Maschi		Femmine	
	18 mesi	adulti	18 mesi	adulti
Altezza al garrese (cm)	65	85	60	80
Altezza alla groppa (cm)	63	78	58	76
Altezza toracica (cm)	34	37	30	35
Larghezza toracica (cm)	25	30	22	28
Larghezza anteriore groppa (cm)	23	25	18	24
Lunghezza tronco (cm)	75	106	65	95
Circonferenza toracica (cm)	80	98	85	94
Peso (kg)	35	65	33	46

Tabella 1: Caratteri biometrici

La capra è una specie a ciclo poliestrale stagionale in quanto presenta cicli estrali continui solo in alcuni mesi dell'anno, intervallati da un periodo di anaestro la cui lunghezza è variabile in funzione della latitudine e della razza (Fatet A. 2011). In determinati climi, le capre sono in grado di riprodursi per tutto l'anno; le razze di provenienza nordica o montana tendono invece ad avere un ciclo riproduttivo basato sulla lunghezza del fotoperiodo. La stagione riproduttiva, per questi animali, inizia quando le giornate cominciano ad accorciarsi, per terminare all'inizio della primavera. Alle nostre latitudini il primo calore si manifesta, solitamente, nei mesi di giugno-luglio; se non vi è accoppiamento si può presentare una lunga

serie di estri ad intervalli regolari di 21 giorni, oppure si può verificare un ulteriore periodo di anaestro. Di norma la ciclicità diventa regolare a partire da agosto-settembre fino alla metà di dicembre.

Nel periodo fertile, le femmine sono ricettive per un periodo che va dalle 2 alle 48 ore, manifestando il loro stato sventolando spesso la coda, vocalizzando più spesso, stando sempre nella vicinanza del becco, a volte perdendo l'appetito e diminuendo la produzione di latte. La gestazione dura in media 150-155 giorni.

I parti generalmente si verificano tra novembre ed aprile con punte massime in gennaio-febbraio.

Il parto, inoltre, coincide con l'inizio della produzione di latte e, per quanto riguarda la razza caprina Girgentana, la quantità di latte prodotto è in media 224 ± 66 L per le primipare e 320 ± 109 L per le pluripare (Di Gerlando R. 2015).

Le caprette raggiungono la pubertà intorno ai 5-6 mesi, con variabilità in funzione delle caratteristiche genetiche ed ambientali. Se non sussistono problemi come stati di carenza nutrizionale, l'accoppiamento può essere effettuato all'età di 7-8 mesi, quando le stesse hanno ormai raggiunto un peso pari al 55-60% di quello adulto.

Ponendo attenzione sulla produzione del latte questa razza è particolarmente vocata all'attitudine lattifera producendo un latte di alta qualità e di elevato valore nutritivo grazie al buon contenuto in protidi, vitamine, lipidi ed elementi minerali, con una percentuale media di grasso e proteine rispettivamente del 4,7% e del 4,2%. Il latte viene destinato al consumo diretto sia per il sapore dolce che per lo scarso odore irascibile dovuto al profilo acidico del latte stesso. Inoltre, questo latte si distingue per la minore dimensione dei globuli lipidici e per la più alta percentuale di acidi grassi a catena media e corta. Queste caratteristiche conferiscono all'alimento una migliore digeribilità, legata al fatto che tali grassi vengono assorbiti rapidamente dalla mucosa intestinale e giungono più velocemente al fegato per la loro metabolizzazione.

1.3 Salvaguardia e valorizzazione della biodiversità zootecnica

La capra, sin dalla sua domesticazione, è stata molto utilizzata dall'uomo per la produzione di diversi prodotti. Oggi l'allevamento di questo animale rappresenta un'importante risorsa economica nei Paesi in via di sviluppo mentre nei Paesi sviluppati questo è in gran parte relegato nelle aree marginali dove può ancora svolgere un ruolo nelle attività agricole a basso impatto (FAO 2007).

Il settore zootecnico siciliano ha mostrato, nell'ultimo decennio, una significativa riduzione degli allevamenti dei piccoli ruminanti soprattutto caprini. Questi hanno sempre rappresentato un'attività idonea a consentire l'utilizzo di aree marginali nelle quali le caratteristiche pedo-morfologiche e climatiche hanno impedito lo sviluppo di un'attività agricola e zootecnica più intensiva. Tali condizioni hanno determinato un ritardo nello sviluppo del comparto primario, il quale ancora oggi incontra notevoli difficoltà nel percorso di trasformazione verso imprese a gestione manageriale.

A tal proposito si è evidenziato che molte razze autoctone al giorno d'oggi hanno uno *status* di estinzione, o addirittura sono in via di estinzione, a causa della loro sostituzione con razze straniere altamente produttive

e soprattutto per il progressivo abbandono di attività rurali a basso reddito. Tutto questo ha determinato l'abbandono dell'allevamento caprino da parte di molti allevatori (Mastrangelo S. 2016). Le razze/popolazioni autoctone in generale sopravvivono in ambienti caratterizzati da condizioni pedoclimatiche e ambientali sfavorevoli alla diffusione di razze cosmopolite e selezionate, dimostrando capacità di adattamento peculiari e interessanti. In altri casi, la sopravvivenza di queste razze è da ricercare in fattori legati alla tradizione e alla cultura contadina locale, che ha garantito la salvaguardia, sino ai giorni nostri, di molte di esse. È solo recentemente che i cambiamenti nel gusto e nella domanda dei consumatori hanno offerto nuove opportunità di rilancio dei prodotti locali e, conseguentemente, di valorizzazione di alcune razze italiane. Ad oggi si può affermare che il recupero del legame indissolubile tra ambiente di allevamento, razza autoctona e prodotto locale costituisce lo strumento più sicuro di conservazione delle Risorse Genetiche Animali (RGA). L'accresciuta consapevolezza degli allevatori e consumatori dell'importanza di conservare la biodiversità è stata accompagnata dall'applicazione, a livello nazionale e regionale, dei vari Regolamenti dell'Unione Europea in materia di conservazione delle RGA. Inoltre, gli incentivi per l'allevamento in purezza di razze autoctone hanno, in molti casi, aiutato la salvaguardia e il rilancio di queste popolazioni; in altri casi, tuttavia, queste misure si sono rivelate inefficaci, dimostrando i limiti di una strategia di conservazione basata esclusivamente sulla scarsa numerosità di una razza (MiPAAF 2012).

In merito alla razza Girgentana si sa che durante il 1983 nella zona costiera, collinare e sub-montana dell'agrigentino la sua consistenza era stimata intorno ai 30.000 capi tuttavia, col subentrare di nuove norme in materia di sanità e il conseguente divieto di stabulazione delle capre entro i centri abitati, si è evidenziata, negli ultimi 20 anni, una riduzione del 90% (Di Gerlando R. 2015).

Questa drastica riduzione è in parte dovuta alla marcata diminuzione del consumo di latte fresco di capra, tant'è che la razza Girgentana è elencata dalla Food and Agriculture Organization con uno *status* di rischio di estinzione (Cañón J. 2006). Tuttavia, per far fronte a questa drastica riduzione si è cercato di mettere a punto determinati programmi di salvaguardia allo scopo di evitarne l'estinzione (ad es. Cornale P. 2014, Parejo J.C. 2015, Vahidi S.M.F. 2014).

Per poterla salvaguardare, si è pensato di utilizzare questa razza negli allevamenti a regime stabulato, inoltre si è anche pensato di utilizzarla per valorizzare terreni aridi e marginali, allevandola in purezza oppure utilizzandola su razze locali con quello che è l'incrocio di sostituzione, al fine di migliorare la morfologia della mammella che deve presentarsi globosa e ben attaccata.

Infine, la conservazione di questa preziosa risorsa genetica è importante per preservare il potenziale futuro allevamento e soprattutto per sviluppare dei sistemi sostenibili di produzione animale. Pertanto, per la gestione efficace delle risorse degli animali da allevamento è molto importante, in particolare, la conoscenza approfondita delle caratteristiche delle razze (Groeneveld L.F. 2010) dato che ciò può rappresentare il punto di partenza per lo sviluppo di un programma di conservazione. A tal fine, le moderne tecnologie molecolari consentono, oggi, la caratterizzazione a livello del DNA delle risorse genetiche autoctone in tempi brevi e ad un costo accessibile.

2 I marcatori molecolari

2.1 Descrizione generale e caratteristiche

Un settore di notevole interesse e di grandi potenzialità è quello basato sull'analisi del genoma mediante rilevazione di marcatori molecolari. Si tratta di un settore che negli ultimi decenni ha visto una rapida evoluzione che si avvantaggia da una parte dello sviluppo di tecnologie sempre più sensibili ed affidabili e dall'altra della conoscenza sempre più approfondita della struttura, organizzazione e funzione degli acidi nucleici.

L'analisi del genoma mediante marcatori molecolari è in grado di rilevare le differenze (polimorfismi) in regioni omologhe tra individui diversi appartenenti alla medesima specie. Le differenze tra individui a livello di sequenza nucleotidica del DNA costituiscono un insieme di marcatori genetici con elevate potenzialità discriminanti e rappresentano un sistema di analisi comparativa dei genomi con un alto livello di dettaglio. L'identificazione e l'isolamento di numerosi enzimi che agiscono sugli acidi nucleici (nucleasi di restrizione, ligasi, polimerasi, chinasi) e la messa a punto di tecniche elettroforetiche di separazione degli acidi nucleici hanno reso possibile quella che è l'analisi della struttura e, in alcuni casi, delle funzioni del materiale ereditario (Figura 4).

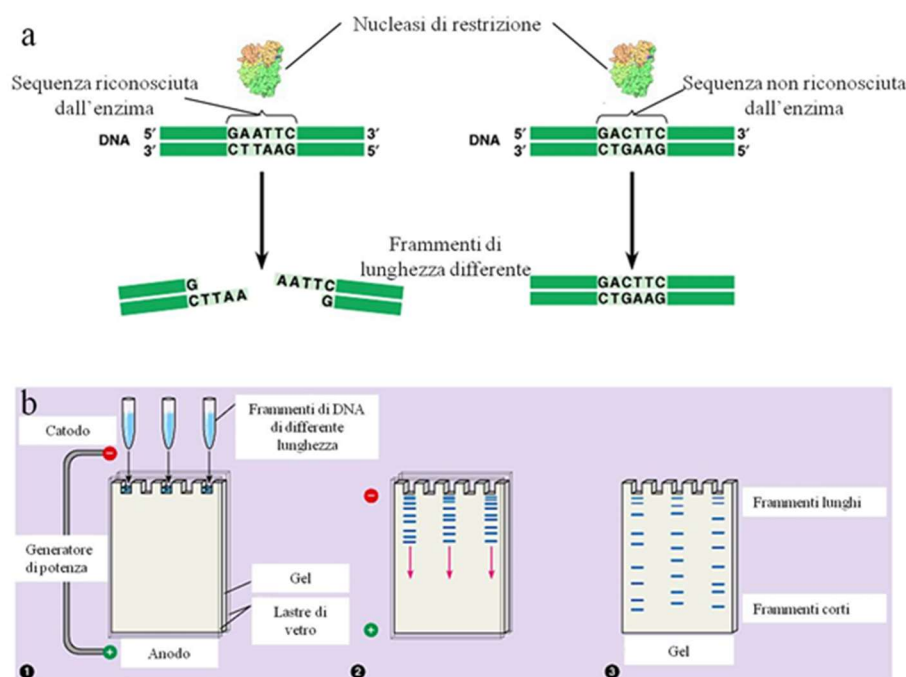


Figura 4: Esempi di tecniche utilizzate per l'analisi strutturale del DNA. a) Frammentazione del DNA con nucleasi di restrizione: un polimorfismo presente all'interno della sequenza riconosciuta da una nucleasi di restrizione impedisce all'enzima di riconoscere il sito di taglio e di generare frammenti di dimensioni inferiori. B) Elettroforesi capillare di frammenti di DNA di lunghezza differente: una popolazione di molecole di DNA di lunghezza differente migra attraverso un gel immerso in un campo elettromagnetico; ciascun frammento percorre una distanza inversamente proporzionale alla sua lunghezza.

Un marcatore molecolare può essere definito come quel *locus* genomico, rilevabile con sonde o inneschi specifici che, in virtù della sua presenza, contraddistingue in maniera caratteristica ed inequivocabile il tratto cromosomico con cui si identifica. Pertanto, i marcatori molecolari non sono generalmente riferibili all'attività di specifici geni, ma si basano direttamente sulla rilevazione di differenze (polimorfismi) nella sequenza nucleotidica del DNA che costituisce il patrimonio ereditario di ciascun individuo. Tali polimorfismi possono essere dovuti ad inserzioni, delezioni, traslocazioni, duplicazioni, mutazioni puntiformi, ecc.

I marcatori molecolari possono essere costituiti da corte sequenze di DNA, come la sequenza di nucleotidi che circonda un polimorfismo a singolo nucleotide (*Single Nucleotide Polymorphism* o SNP), o da lunghe sequenze di DNA come i microsatelliti.

I marcatori molecolari più frequentemente utilizzati sono:

- RFLP (*Restriction Fragment Length Polymorphism*);
- AFLP (*Amplified Fragment Length Polymorphism*);
- RAPD (*Random Amplified Polymorphic DNA*);
- SNP (*Single Nucleotide Polymorphism*);
- Microsatelliti (*Short Tandem Repeats* o STRs).

I marcatori molecolari possono essere usati, per esempio, per studiare la relazione tra la trasmissione di una malattia ereditaria e le sue cause genetiche (una particolare mutazione di un gene che codifica una proteina difettosa) (Poetsch M. 2004, Pelotti S. 2007). Una delle caratteristiche peculiari più importanti dei marcatori è infatti l'ereditabilità, ossia la possibilità di essere trasmessi da una generazione alla successiva. Altre caratteristiche distintive sono l'associazione ad uno specifico *locus*, l'elevato polimorfismo, la possibilità di essere applicati universalmente e di essere distribuiti uniformemente nel genoma.

2.2 Single Nucleotide Polymorphisms o SNPs

Tra i marcatori sopra citati, quelli che negli ultimi tempi hanno preso il sopravvento sono gli SNPs (*Single nucleotide polymorphisms*). Questo tipo di marcatore consente di mettere in evidenza il polimorfismo riconducibile a differenze di singoli nucleotidi (Figura 5) (Jordan S. A. 1994). Marcatori SNPs possono essere messi in evidenza sia a carico delle regioni trascritte di DNA che di quelle non trascritte.

Data la frequenza con cui avvengono le mutazioni puntiformi, gli SNPs sono considerati i marcatori con il più alto potenziale di polimorfismo genomico rilevabile. Gli SNPs possono presentarsi all'interno di una sequenza codificante di un gene, all'interno di una regione intronica o in una regione intergenica. Gli SNPs all'interno di un gene, in ogni caso, non necessariamente modificano la sequenza amminoacidica codificata, dal momento che il codice genetico è degenerato. Uno SNP che genera in tutte le sue forme lo stesso peptide è detto sinonimo (*synonymous*); in caso contrario è detto non-sinonimo (*non-synonymous*). Gli SNPs che non si trovano in una sequenza codificante possono, in ogni caso, presentare sequenze negative sullo *splicing* o sul legame dei fattori di trascrizione.

In generale, gli SNPs costituiscono il 90% di tutte le variazioni genetiche. SNPs con frequenza dell'allele minore (*Minor Allele Frequency* o MAF) pari o maggiore all'1% sono presenti ogni circa 100-300 paia di basi lungo l'intero genoma.

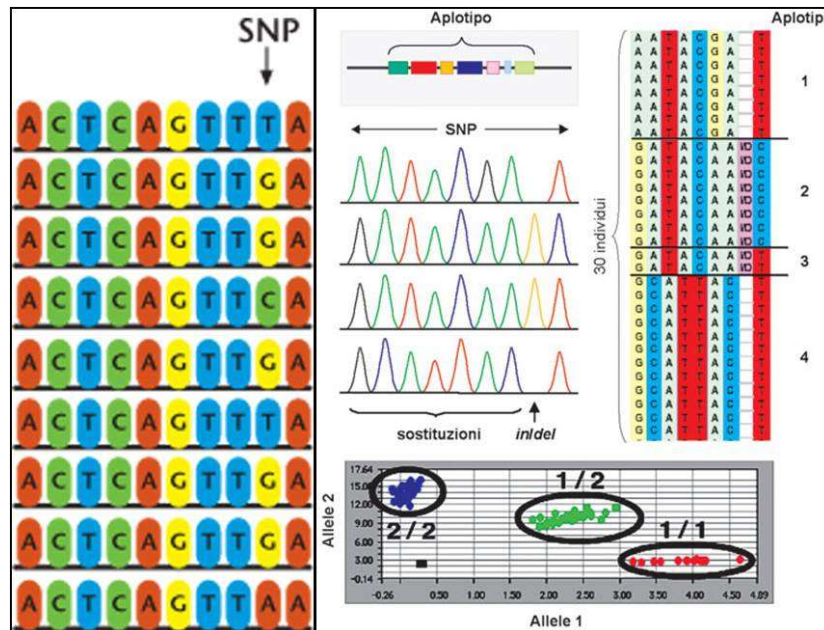


Figura 5: Rappresentazione schematica di SNP. A sinistra: allineamento multiplo di sequenze che mostrano polimorfismo dovuto alla sostituzione di singoli nucleotidi. A destra: una rappresentazione grafica di un aplotipo molecolare basato su marcatori SNPs.

Gli SNPs si possono identificare tramite sequenziamento di tipo Sanger, una volta identificati per poterli evidenziare si può applicare una semplice PCR con base specifica per lo SNP in posizione 5'. Il sequenziamento dei prodotti di amplificazione e l'allineamento delle sequenze mette in evidenza quelle che sono le differenze dovute a mutazioni puntiformi (inserzioni, delezioni, sostituzioni). Un altro metodo utile per individuare l'informazione contenuta negli SNPs, ossia il loro genotipo, è attraverso l'utilizzo di *microarrays*, che permettono l'analisi simultanea di decine di campioni per ricavare l'informazione dei genotipi di centinaia di migliaia di SNPs distribuiti uniformemente sull'intero genoma.

A tal fine i marcatori SNPs vengono infatti considerati dagli esperti del settore come la classe di marcatori molecolari del futuro. È già stata accertata l'idoneità di tali marcatori in analisi filogenetiche ed evolutive, dovuta soprattutto alla loro capacità di mettere in evidenza aplotipi molecolari altamente conservati.

2.3 Principali utilizzi dei marcatori SNPs nel settore zootecnico

Come sopracitato, gli SNPs sono marcatori che consentono di mettere in evidenza il polimorfismo riconducibile a differenze di singoli nucleotidi. Questa classe di marcatori, grazie al continuo sviluppo della tecnologia applicata alla biologia molecolare, si avvale oggi, di tecniche ad alta efficienza, facilmente automatizzabili e sempre meno costose, ciò costituisce uno strumento molto promettente su cui basare ad esempio molti sistemi di tracciabilità efficienti (Negrini R. 2008).

Questo tipo di tecnologia può essere d'aiuto soprattutto nel settore agroalimentare, in particolare nel campo della tracciabilità e della sicurezza alimentare, infatti i metodi basati sull'analisi del DNA, possono aiutare nell'identificazione della specie degli animali che sono presenti in una derrata alimentare.

In merito all'utilizzo dei marcatori per la tracciabilità e per la sicurezza alimentare nel corso degli anni, per salvaguardare alcuni prodotti con marchi europei (DOP e IGP), sono stati creati dei pannelli di marcatori SNPs in grado di tracciarli. Un esempio è il pannello di marcatori disponibile per il Vitellone Bianco dell'Appennino Centrale IGP ottenuto dalle razze da carne Chianina, Marchigiana e Romagnola (Negrini R. 2008). In un altro studio (Heaton M.P. 2005) è stato selezionato un set di 20 marcatori SNPs bovini per verificare l'accuratezza della tracciabilità negli impianti di macellazione di bovini da carne o da latte. Pertanto, si è visto che questi marcatori possono facilitare i programmi di tracciabilità ad alto rendimento.

I marcatori SNPs vengono anche utilizzati per studi di diversità genetica, per esempio, proprio per la specie caprina, Nicoloso e coll. (Nicoloso L. 2015) hanno utilizzato un chip di SNPs a media densità per valutare la diversità genetica delle popolazioni caprine italiane. Grazie ai risultati descritti da questi autori, si è visto che l'attuale diversità genetica delle popolazioni caprine italiane è dovuta ad effetti combinati di deriva genetica, presenza o mancanza di flusso genico e, in una certa misura, ai sistemi di gestione tradizionale e alla recente storia demografica.

Le informazioni derivanti dagli SNP *arrays* possono essere utilizzate per condurre studi di associazione *genome-wide* ed in particolare per incrementare le conoscenze sui caratteri complessi aumentando l'efficienza di selezione negli animali di interesse zootecnico utilizzando, in particolar modo, la selezione assistita da marcatori e la selezione genomica. Per ottimizzare uno studio di associazione *genome-wide* è importante conoscere il *linkage disequilibrium* (LD) tra marcatori al fine di stabilire il numero di marcatori necessari per la selezione genomica e per mappare alcune malattie genetiche (Gurgul A. 2014). Gli SNP *arrays* vengono anche utilizzati per identificare le *runs of homozygosity* (ROHs). Queste ROHs sono regioni contigue di genotipi omozigoti dove i due aplotipi ereditati dai genitori sono identici. Ciò si traduce nella formazione di ROHs con lunghezze diverse: i segmenti più lunghi rappresentano *inbreeding* recente e quelli più corti sono associati ad *inbreeding* proveniente da antenati lontani. La lunghezza e la frequenza delle ROHs possono dare informazioni sulle ascendenze e sulla storia delle razze/popolazioni (Purfield D.C. 2012). Come è noto, questi sono solo alcuni degli utilizzi degli SNPs e delle loro possibili applicazioni.

3 Le Copy Number Variants (CNVs)

3.1 Definizione ed identificazione tramite diverse metodologie

Negli ultimi anni, molte ricerche si sono focalizzate sulle variazioni del numero di copie o *Copy Number Variants* (CNVs) che rappresentano una variazione strutturale del genoma e sono considerate una componente importante di diversità genetica (Redon R. 2006) così come i polimorfismi a singolo nucleotide (SNPs). Redon e coll. (Redon R. 2006) hanno definito una CNV come un segmento di DNA, di dimensioni variabili da 1 kilobase (kb) a qualche megabase (Mb), presente in un numero di copie variabile rispetto a un genoma di riferimento. Le regioni di CNVs possono riguardare geni attivi o gruppi di geni, come promotori, *enhancer* o altre sequenze funzionalmente importanti (Gurgul A. 2014). Le CNVs possono insorgere a causa di diversi meccanismi molecolari quali la duplicazione, l'inserzione o la delezione di frammenti di DNA, la cosiddetta *non-allelic homologous recombination* (NAHR), oppure la *non homologous end joining* (NHEJ), lo *slippage* di replicazione e la retrotrasposizione. Il meccanismo più comune presente negli esseri umani è la NAHR ed il meno comune è la retrotrasposizione (Gurgul A. 2014).

Le CNVs sono presenti in una vasta gamma di organismi, non solo negli esseri umani (Redon R. 2006, Sebat J. 2004, McCarroll S.A. 2006, Conrad D.F. 2006) ma anche nelle specie animali quali topi (Graubert T.A. 2007, She X. 2008), scimpanzé (Perry G.H. 2008), macaco rhesus (Lee A.S. 2008), bovini (Liu G.E. 2010), canidi (Chen W.K. 2009), avicoli (Griffin D.K. 2008), moscerini della frutta (Emerson J.J. 2008), *Caenorhabditis elegans* (Maydan J.S. 2010), così come nelle specie vegetali, come il mais (Springer N.M. 2009), l'*Arabidopsis thaliana* (Ossowski S. 2008) e persino i funghi, come nel *Saccharomyces cerevisiae* (Carreto L. 2008). Le CNVs possono anche variare tra individui all'interno di una stessa specie (Schridder D.R. 2010). Le CNVs possono insorgere *de novo* in un organismo e in questo caso si parlerà di CNV somatica o possono anche essere il risultato di interruzioni nel processo di ricombinazione nelle cellule germinali, ed in quest'ultimo caso saranno ereditabili. Tuttavia, la presenza delle CNVs potrebbe non essere neutrale per un organismo. Numerosi studi hanno dimostrato che queste variazioni influenzano le caratteristiche fenotipiche, le complesse basi del comportamento, la suscettibilità/resistenza alle malattie, nonché la comparsa di disturbi genetici (Buckland P.R. 2003, Fanciulli M. 2007, Schaschl H. 2009). Le CNVs possono riguardare un singolo gene o un insieme di geni adiacenti così come possono riguardare le mutazioni di alleli nel caso di malattie complesse (Henrichsen C.N. 2009).

Le CNVs possono alterare le sequenze che regolano l'espressione genica, come gli *enhancer* (McC061), (Nguyen D.Q. 2006) o i promotori, così come possono portare all'esposizione di alleli recessivi (quando si verifica una delezione del gene dominante) o persino all'inattivazione di alcuni geni (quando avviene una delezione all'interno di un gene). Pertanto, si è visto che alcuni fenotipi possono derivare non dai

cambiamenti del numero di copie di una data CNV, ma da un'alterazione strutturale in un frammento del genoma, che comporta l'interruzione di una via metabolica, indipendentemente dal dosaggio del gene (Henrichsen C.N. 2009). Infine, l'interazione con fattori genetici o ambientali può ulteriormente influenzare quello che è l'effetto fenotipico delle CNVs.

Per quanto riguarda le specie zootecniche, negli ultimi anni, la ricerca ha portato notevoli risultati. Prima di tutto, la costruzione di mappe a bassa risoluzione di CNVs per bovini, equini, caprini, ovini, suini, canidi, polli, anatre e tacchini, ha fornito una panoramica dei loro genomi e ha dimostrato che queste variazioni sono ampiamente diffuse in queste specie. Inoltre, come per la specie umana, le CNVs sono state associate a diversi fenotipi e alla suscettibilità ad alcune malattie, oltre a disturbi dello sviluppo, quali, ad esempio, le diverse pigmentazioni (colore bianco del mantello negli equini, nei suini e negli ovini) ed alcuni caratteri morfologici (piumaggio tardivo e la pea comb nel pollo), osteopetrosi, displasia ectodermica anidrotica, tossicosi del rame, intersessualità e degenerazione del cono retinico (Clop A. 2012).

Le CNVs possono essere identificate con l'uso di un'ampia gamma di tecniche quali FISH (*Fluorescent In Situ Hybridization*), CGH (*Comparative Genomic Hybridization*), aCGH (*Comparative Genomic Hybridisation array*), *Southern blotting*, PFGE (*Pulsed-Field Gel Electrophoresis*), MAPH (*Multiplex Amplifiable Probe Hybridization*), MLPA (*Multiplex Ligation-dependent Probe Amplification*), PRT (*Parologue Ratio Test*) e qPCR (*quantitative Polymerase Chain Reaction*).

Inoltre, le metodologie molecolari all'avanguardia e le tecniche computazionali avanzate consentono l'identificazione delle CNVs su scala genomica tramite tecnologie *high-throughput* NGS (*Next-Generation Sequencing*) e *SNP genotyping arrays*.

Per identificare le CNVs dai dati provenienti da un *array*, è possibile utilizzare la combinazione di due misure delle intensità del segnale: LRR (*Log R Ratio*) e BAF (*B Allele Frequency*). Una deviazione significativa dalla distribuzione prevista di questi due parametri implica un numero diverso di copie di un dato allele (Wang K. 2007).

Ad oggi, molti studi riportano l'utilizzo degli *SNP genotyping arrays* per l'identificazione della CNVs. In particolare, si utilizzano i *Genotyping BeadChip* Illumina che permettono di eseguire l'analisi di genotipizzazione su migliaia di loci. L'elemento essenziale della tecnologia *BeadChip* è l'attacco degli oligonucleotidi a microsfele di silice di 2 micron che si autoassemblano nei micropozzetti. Le sfere vengono poi depositate casualmente in pozzetti su un substrato (ad esempio, un vetrino). Ciascuna microsfera è ricoperta da centinaia di migliaia di copie di uno specifico oligonucleotide che agisce come "sequenza di cattura" (Figura 6).

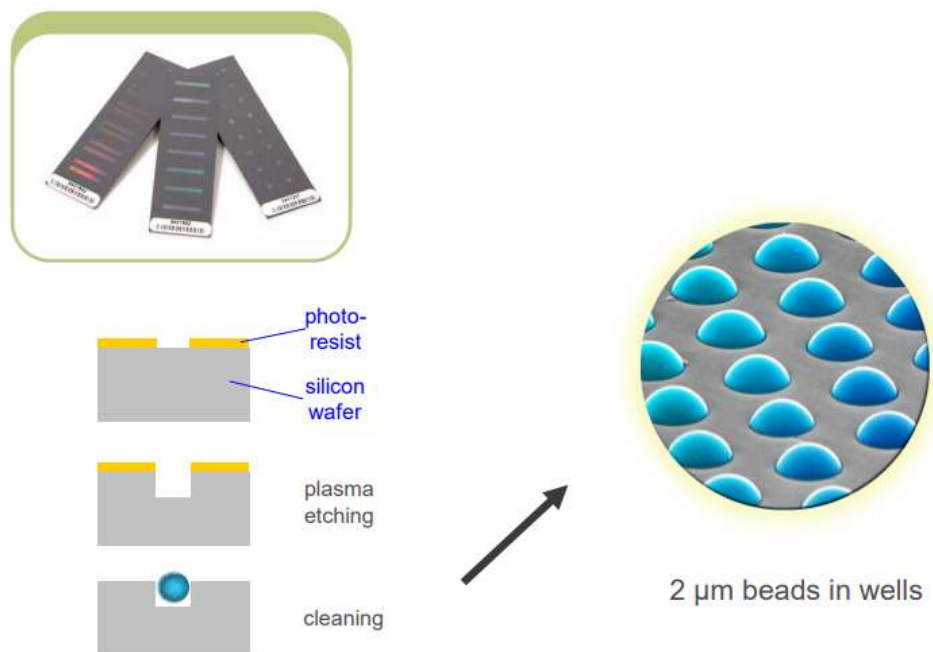


Figura 6: microproduzione di pozzetti BeadChip

Il protocollo prevede la preparazione del campione e l'ibridazione del campione di DNA non marcato ad una biglia legata ad un oligomero di 50 pb, e la successiva incorporazione di una singola base marcata durante l'estensione del filamento. La fase finale una volta che il *BeadChip* è pronto, è la lettura dello stesso da parte dello strumento HiScanSQ che altro non è che uno scanner. Le intensità di fluorescenza di ogni SNP vengono registrate dallo strumento e rese disponibili all'operatore grazie al software GenomeStudio 2.0.

3.2 Importanza delle CNVs nelle specie di interesse zootecnico

3.2.1 CNVs nei bovini

La scoperta delle *Copy Number Variants* (CNVs) ha portato i ricercatori a pensare che ci siano più varianti del DNA genomico di quanto inizialmente creduto, infatti le CNVs rappresentano una fonte importante di diversità genetica nell'uomo, così come nelle specie di interesse zootecnico, dimostrando di essere associate a fenotipi di rilevanza economica ed alla resistenza ad alcune malattie. Possono, inoltre, spiegare parte della diversità fenotipica e dell'adattamento alle condizioni ambientali.

A tal proposito nel corso degli anni sono stati fatti diversi studi, in particolare sui bovini, di cui si possono trovare una moltitudine di articoli scientifici. Nei primi studi, grazie anche ai paragoni con altri mammiferi come l'uomo, si è visto che molte CNVs andavano ad arricchire i geni che erano correlati alla risposta ambientale e non solo (Fadista J. 2010). È stato visto che molte erano correlate con geni responsabili della risposta immunitaria, molti dei quali sono importanti per la detossificazione quando si è in presenza di

xenobiotici. Per l'immunità innata e adattativa difatti queste famiglie di geni comprendono i trasportatori degli ATP-binding cassette (ABC) e il citocromo P450 come dimostrato nel lavoro di Yang e coll. in cui si notava una differente espressione dei livelli del CYP4A11 per le differenti CNVs (Yang M. 2017), le beta-difensine, i loci dei recettori delle cellule T ed il *Major Histocompatibility Complex* (MHC) bovino (Liu G.E. 2010, (Hou Y. 2011). Molte CNVs sono presenti all'interno di geni responsabili di processi biologici che stimolano la risposta difensiva da parte della cellula intervenendo quindi nella risposta immunitaria (Sasaki S. 2016). Alcune CNVs sono state trovate frequentemente nei geni che codificano per i recettori olfattivi come i sensori dell'odore (Di Gerlando R. 2018, de Lemos M.V.A. 2018). Sono state identificate CNVs in geni responsabili di alcune funzioni biologiche come il metabolismo di amminoacidi, la segnalazione e l'interazione cellula-cellula, la funzione e sviluppo del sistema nervoso, il trasporto molecolare e i cicli cellulari; altre sono risultate coinvolte nel *signaling* di *pathways* biologici come quello riguardante l'ephrina B e la degradazione della betaina di glicina, ma anche nella disfunzione mitocondriale e nella fosforilazione ossidativa (Huang J. 2017). È stato ipotizzato anche che, in base alla differenza nella frequenza delle CNVs tra razze bovine, alcune possano sorgere indipendentemente e contribuire a generare differenze rilevabili e, quindi, che siano correlate alla formazione e all'adattamento della razza, e associate ai processi di domesticazione (Liu G.E. 2010, Pierce M.D. 2018). Un classico esempio è lo studio fatto sui bovini del Sud Africa da Wang e coll. (Wang M.D. 2015) in cui è stato visto come le differenti CNVs associate a diversi geni potessero spiegare quella che è la diversità fenotipica e l'adattamento alle condizioni ambientali. Nello studio di Bae e coll. (Bae J.S. 2010) si è evidenziato come le CNVs possano portare a scoprire i geni legati ai caratteri della produttività della carne, fornendo quindi delle risorse genomiche aggiuntive necessarie per andare ad analizzare quei fenotipi importanti per l'impatto economico e per selezionare anche quei bovini da portare sulla tavola del consumatore. Sempre per quanto riguarda la produttività, altri lavori hanno mostrato come le CNVs abbiano un potenziale effetto sui caratteri della produzione del latte ma anche sul contenuto di proteine e di grassi (Jiang L. 2013, Seroussi E. 2010). È stato visto anche, sempre su caratteri importanti a livello economico, come le CNVs siano presenti in geni coinvolti in processi metabolici e cellulari, nella regolazione biologica, nella risposta agli stimoli esterni, nella riproduzione e nella crescita (da Silva J.M. 2016). In quest'ultimo caso, come mostrato nello studio di Prinsen e coll. (Prinsen R.T.M.M. 2016), i CNVs identificati erano coinvolti nello sviluppo e nella funzione del sistema scheletrico e muscolare, nel metabolismo dei lipidi e nello sviluppo embrionale (Huang J. 2017). Alcune CNVs sono presenti anche nei QTL associati alla qualità della carcassa, alla riproduzione, alla crescita, alla produzione di latte, alla salute, alle misure corporee ed al peso dell'animale (Huang J. 2017). In studi di associazione *genome-wide* (Prinsen R.T.M.M. 2017), alcune CNVs sono state associate a caratteri legati alla produzione di latte e a caratteri riguardanti le infezioni intramammarie, fornendo quindi un punto di partenza per il miglioramento della produzione di latte e contemporaneamente dello stato di salute degli animali.

3.2.2 CNVs negli ovini

Nel corso degli anni sono state fatte ricerche per rilevare la presenza di CNVs anche nella specie ovina. I primi studi hanno previsto dei paragoni con quanto trovato nei bovini, data l'elevata omologia tra le due specie ruminanti. Le CNVs possono, infatti, interessare il cosiddetto locus *Agouti* che contribuisce alla colorazione del mantello (Fontanesi L. 2009, Norris B.J. 2008). In uno studio sul metabolismo lipidico negli ovini, ovvero quel processo chiave che determina la composizione di acidi grassi nel latte, si è visto che è difficile valutare come le CNVs siano coinvolte nella variabilità di questi caratteri (Fontanesi L. 2011). Come visto per i bovini, anche negli ovini si è riscontrata una interazione tra CNVs e l'attività dei recettori olfattivi, che contribuiscono ad una serie di molteplici attività, come la percezione degli odori, degli stimoli chimici, il riconoscimento delle fonti alimentari e della prole (Liu J. 2013). Nella specie ovina, così come in quella bovina, le CNVs si trovano all'interno di geni che sono coinvolti in numerosi processi biologici, nelle funzioni molecolari e nelle componenti cellulari (Yan J. 2017). Sono state identificate CNVs coinvolte nella risposta ambientale e presenti, per esempio, nella famiglia di geni *Zinc Finger* (ZNF), come ZNF622 e ZNF688, coinvolti nel regolare caratteri biologici evolutivamente divergenti (Ma Y. 2015). Alcune CNVs correlate a caratteri fenotipici sono state trovate in geni associati a fenotipi divergenti come il caso della presenza/assenza della coda grassa in alcune razze ovine presenti nello studio di Zhu e coll. (Zhu C. 2016). In conclusione, si può affermare che i risultati ottenuti nel corso degli anni su questa specie forniscono delle preziose informazioni a supporto delle indagini sulle variazioni genomiche strutturali presenti nel genoma ovino, nonostante si sia evidenziato come i differenti metodi utilizzati per l'individuazione delle CNVs possano influenzare i risultati ottenuti (Jenkins G.M. 2016).

3.2.3 CNVs nei caprini

Per la specie caprina, gli studi che riportano l'identificazione di CNVs sono in numero limitato considerando anche il limitato numero di studi eseguiti a livello genomico su questa specie (Fontanesi L. 2010). I primi lavori di identificazione di CNVs caprine sono stati ottenuti utilizzando l'aCGH in cui erano presenti il genoma bovino e quello caprino, per identificare CNVs interspecifiche ricorrenti tra queste due specie di ruminanti. Grazie a queste tecnologie si sono potute evidenziare CNVs importanti a livello fenotipico, tra cui quelle presenti nel locus *Agouti* (Fontanesi L. 2009) importante per la colorazione del mantello. Inoltre, sempre attraverso questi studi di ibridazione comparativa, sono state trovate CNVs presenti in geni con funzioni ambientali (Fontanesi L. 2010). Grazie ai diversi studi condotti, le CNVs sono state evidenziate in geni a rapida evoluzione tra cui il gene *ASIP*, responsabile del colore del mantello, ed alcuni geni associati ai caratteri comportamentali, alla risposta immunitaria e ai caratteri produttivi che hanno contribuito alla formazione delle diverse razze caprine soggette a domesticazione (Dong Y. 2015).

3.3 Validazione delle CNVs con differenti metodi

Uno dei passaggi più importanti negli studi di identificazione di CNVs è la validazione dei risultati ottenuti utilizzando una metodologia diversa da quella usata per identificarle. Il metodo più comunemente utilizzato è la Real-Time PCR, anche nota come PCR quantitativa in tempo reale, mentre di più recente utilizzo sono i dati provenienti da *whole-genome sequencing* (WGS) grazie alle tecnologie NGS che ormai hanno raggiunto costi contenuti e che vengono preferite per l'enorme quantità di dati generati.

Per quanto riguarda la Real-Time PCR, questa è in grado di amplificare e quantificare il DNA in maniera simultanea ovvero misura l'amplificazione in tempo reale durante la fase esponenziale della PCR, permettendo di ottenere risultati molto più accurati rispetto alla PCR tradizionale. La quantificazione viene effettuata determinando un segnale di fluorescenza generato in seguito al legame di un reagente al prodotto amplificato nei primi stadi della reazione di PCR. Nello specifico, il metodo consiste nell'uso di sonde gene-specifiche (sonde TaqMan) che si appaiano nella zona compresa fra i due *primers* (*forward* e *reverse*) e contenenti una molecola fluorescente detta *reporter* all'estremità 5' ed un'altra molecola della *quencher* all'estremità 3'. L'accumularsi del prodotto amplificato viene rivelato monitorando l'incremento di fluorescenza del *reporter* (Figura 6).

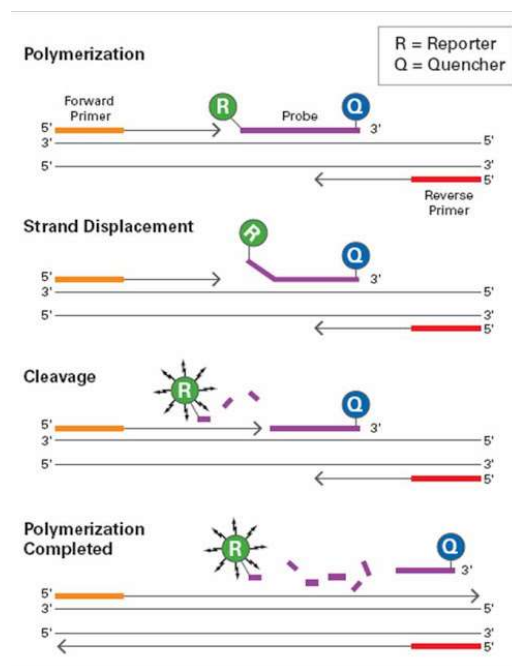


Figura 7: processo di reazione delle sonde TaqMan.

Tra i vantaggi di questa metodica si hanno la velocità, la sensibilità, la variabilità ridotta, la possibilità di effettuare reazioni multiple e quantificazioni relative o assolute.

La Real-Time PCR è la tecnica più utilizzata per la validazione delle CNVs soprattutto per la sensibilità del metodo e l'affidabilità dei risultati. La validazione viene effettuata su un numero limitato di CNVs e/o di *CNV regions* (CNVRs) ovvero regioni che contengono uno o più CNVs consecutivi/sovrapposti. Negli studi in cui questa tecnica viene adottata, vengono preliminarmente scelte un numero ridotto di CNVs e/o

CNVRs da validare (Sasaki S. 2016, Jiang L. 2013, Prinsen R.T.M.M. 2016) rispetto a quelle identificate. A titolo di esempio, nel lavoro di Jiang e coll. (Jiang L. 2013), tra le 368 CNVRs identificate ne sono state scelte 17, che corrispondono al 4,6% del totale, delle quali 13 sono state confermate con successo. Per cui, nonostante gli autori riportino nel lavoro che il 76,5% delle CNVRs è stato validato con successo, questa percentuale si riferisce esclusivamente alle 17 CNVRs scelte, mentre se rapportata al totale, il risultato equivale a dire che il 3,5% di tutte le CNVRs identificate era realmente presente.

Un altro modo per poter validare le CNVs è attraverso l'utilizzo dei dati provenienti da WGS ovvero attraverso il sequenziamento dell'intero genoma di alcuni individui. Le tecnologie NGS nel corso degli anni si sono evolute rapidamente tant'è vero che uno dei primi sistemi NGS, il 454 GenomeSequencer FLX Instrument (Roche Applied Science), produceva 25 milioni di basi sequenziate per volta (Margulies M. 2005), mentre sistemi più recenti riescono a produrre 18.000 miliardi di basi, cioè circa 1 milione di volte in più. Queste nuove tecniche hanno il vantaggio di ridurre costi e tempi, ma soprattutto di ottenere grandi quantità di informazioni con un unico ciclo di sequenziamento.

In questo caso, per la validazione delle CNVs, gli individui sequenziati non sono in numero elevato ma le informazioni ricavabili da dati di *sequencing* sono enormemente superiori. In questo modo, molte delle CNVs/CNVRs identificate con i dati dei BeadChip potranno essere confrontate con quelle identificate con dati NGS (da Silva J.M. 2016, Dong Y. 2015) sfruttando la maggiore risoluzione propria di questa tecnologia.

Capitolo II: Parte Sperimentale

1 Obiettivo della tesi

Nel corso degli ultimi anni, molti studi nel settore zootecnico si sono focalizzati sulle CNVs in quanto variazioni strutturali del genoma che, al pari degli SNPs, rappresentano una componente importante di diversità genetica così come di variabilità fenotipica e di suscettibilità/resistenza ad alcune malattie. Pertanto l'obiettivo della presente tesi è stato quello di identificare le CNVs presenti nel genoma della razza caprina Girgentana tramite l'utilizzo dei dati di genotipizzazione provenienti dai GoatSNP50 BeadChip Illumina, contenenti 53.347 SNPs. Nello specifico, l'identificazione delle CNVs è stata effettuata con l'utilizzo di due differenti software, il PennCNV e l'SVS, ed i risultati sono stati validati tramite sequenziamento individuale con tecnologie high-throughput Next-Generation Sequencing.

2 Materiale e metodi

2.1 Genotipizzazione di capre di razza Girgentana con BeadChip Illumina ed analisi bioinformatiche tramite software PennCNV e SVS

2.1.1 Campionamento e preparazione del campione

Per l'identificazione delle CNVs nella razza caprina Girgentana, sono stati utilizzati 36 campioni provenienti da 9 allevamenti situati sul territorio siciliano.

Il DNA genomico è stato ottenuto da campioni di sangue prelevati dalla vena giugulare utilizzando provette Vacutainer da 6 ml contenenti EDTA come anticoagulante. Il protocollo di estrazione *salting-out* utilizzato prevede l'isolamento dal sangue intero di un *pellet* di globuli bianchi attraverso ripetuti lavaggi utilizzando una soluzione di lisi dei globuli rossi ed una soluzione di salina (Tris-buffered saline, TBS) a pH 7,4. Il *pellet* ottenuto è quindi risospeso in una soluzione 1X TE a pH 8,0 a cui viene aggiunta una soluzione contenente Proteinasi K, 0,5M EDTA e 10% SDS. Successivamente, i campioni vengono lasciati in un bagno termostatico over-night a 50-54°C. Il giorno successivo, dopo l'aggiunta di una soluzione satura di NaCl 6M, i campioni sono stati centrifugati per consentire la precipitazione delle proteine. Il surnatante viene trasferito in una nuova provetta da 15 ml contenente etanolo assoluto freddo per consentire la flocculazione del DNA. Il "fiocco" di DNA viene recuperato e trasferito in un tubo da centrifuga da 1,5 ml per essere facilmente lavato con etanolo freddo al 70%. Infine, l'etanolo viene rimosso ed il *pellet* viene risospeso in acqua ultrapura e conservato a 4°C fino al momento della quantificazione.

Per misurare la concentrazione di DNA e valutarne la qualità viene utilizzato uno spettrofotometro NanoDrop ND1000 (NanoDrop Technologies, Inc., Wilmington, DE, USA).

Una volta quantificati, i campioni sono stati diluiti a una concentrazione finale di 50 ng/μL, concentrazione ottimale per il protocollo Infinium di Illumina. Il protocollo Infinium di Illumina sta alla base del sistema di genotipizzazione Goat SNP50K BeadChip array che consente di analizzare 53.347 SNPs per singolo campione di DNA. Il saggio Infinium si basa sull'ibridazione diretta di *targets* genomici sulle sequenze associate all'*array*. In particolare, il protocollo prevede un'amplificazione del DNA durante il primo giorno di lavoro, la preparazione dei BeadChip, e quindi l'ibridazione del DNA amplificato sul *genotyping array* nel secondo giorno e, infine, nel terzo giorno di lavoro l'estensione della singola base che è seguita dalla colorazione fluorescente e dalla scansione del BeadChip con l'HiScan Illumina (Figura 8). Le intensità di fluorescenza di ogni singolo SNP vengono registrate dall'HiScan e rese disponibili in vari formati mediante il software GenomeStudio.

Dato l'elevato numero di SNPs che possono essere analizzati sui chip Infinium, la semplicità della preparazione dei campioni e la relativa facilità di analisi dei dati, questa piattaforma è stata ampiamente utilizzata per importanti studi in numerose specie.

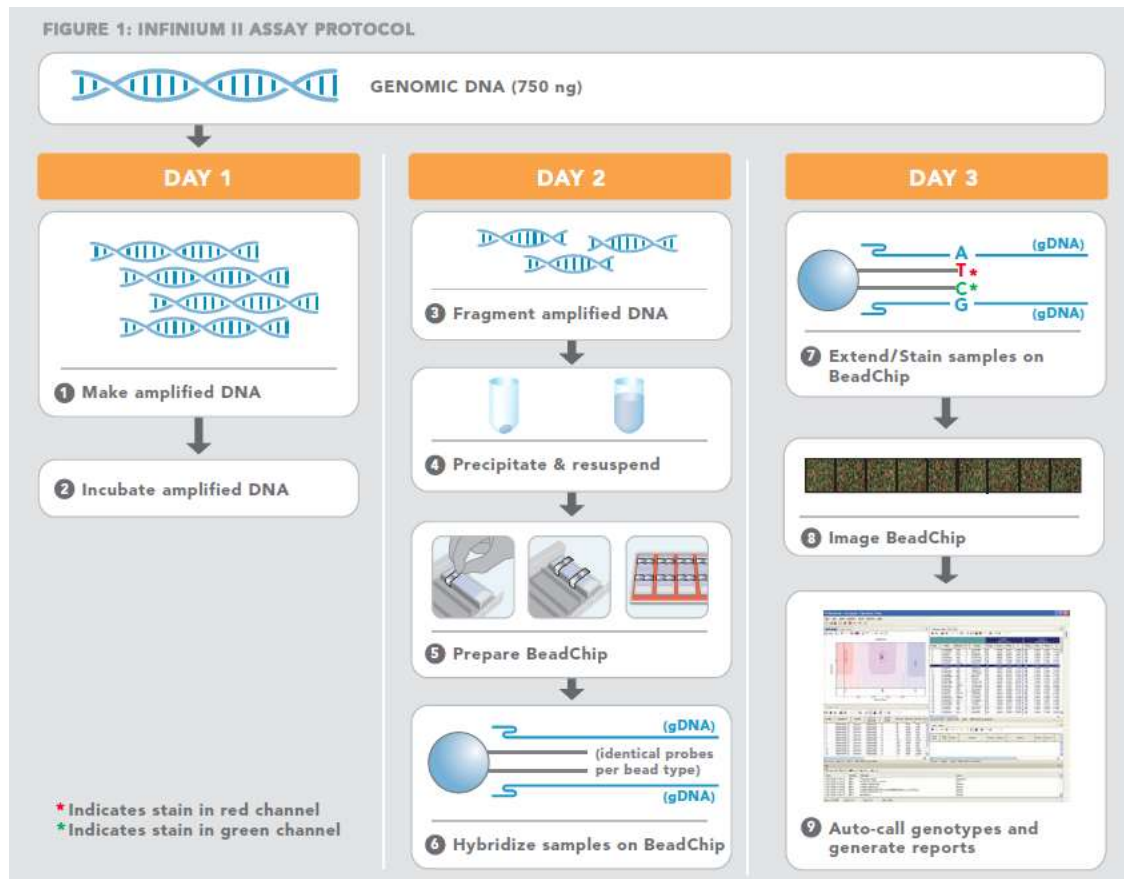


Figura 8: protocollo Infinium Illumina

2.1.2 I software GenomeStudio 2.0, PennCNV e SVS 8.7.0

Una volta terminata la lettura dei BeadChip, grazie al software GenomeStudio 2.0, è possibile esportare i dati da analizzare con i software PennCNV e SVS v8.7.0.

I files di input per entrambi questi software contengono i valori di Log R Ratio (LRR) e B Allele Frequency (BAF) per ogni SNP e per ogni individuo. Per le successive analisi di identificazione delle CNVs, dal dataset originale contenente 53.347 SNPs sono stati eliminati gli SNPs posizionati sui cromosomi sessuali e quelli non mappati. Il dataset finale, per entrambi i software, conteneva 50.619 SNPs e 36 individui.

Il software PennCNV (<http://penncnv.openbioinformatics.org/en/latest/>) utilizza l'algoritmo HMM (Hidden Markov Model) per rilevare CNVs con un'alta risoluzione utilizzando i dati provenienti dall'Illumina Infinium assay. La combinazione di LRR e BAF può essere utilizzata insieme per determinare i diversi numeri di copie e per differenziare le regioni dette LOH (Loss Of Heterozygosity) rispetto alle regioni normali. A parte i files contenenti i valori di LRR e BAF relativi ai singoli animali oggetto dello studio, il

software PennCNV richiede il file del BAF della popolazione di riferimento per ogni SNP ed il file della distanza tra SNP adiacenti, utile per determinare la probabilità di avere una variazione dello stato del numero di copie tra di essi (Wang K. 2007). Il file del BAF della popolazione viene calcolato direttamente da PennCNV sulla base del valore di BAF di ogni SNP in ogni individuo. PennCNV utilizza anche un file con il contenuto in Guanine-Citosine (GC) calcolato in finestre da 1 Mb attorno ad ogni SNP (500 kb a destra e 500 kb a sinistra). Questo file viene utilizzato per aggiustare le cosiddette *genomic waves*. Infine, la qualità del dataset finale viene valutata applicando i seguenti filtri: deviazione standard di LRR, $LRR_SD < 0,30$; BAF drift $< 0,01$ e *waviness factor* (WF) $-0,05 > WF > 0,05$ (Figura 9).

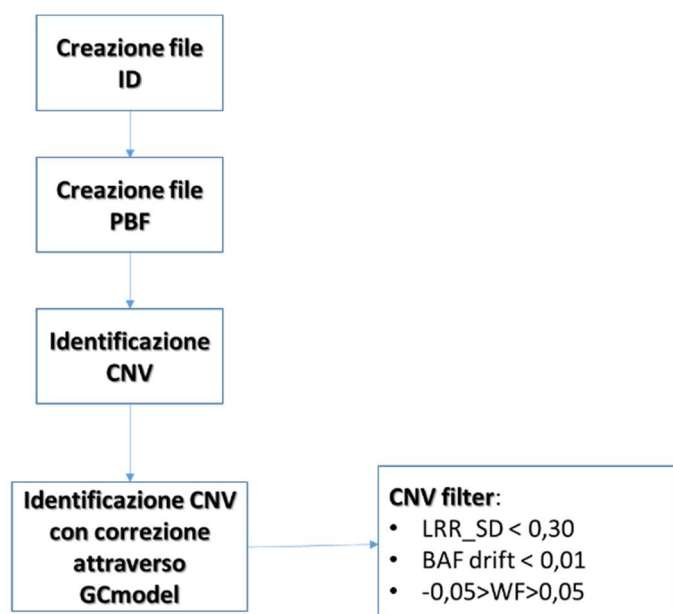


Figura 9: flusso di lavoro PennCNV

Altro software per l'identificazione delle CNVs è SVS 8.7.0, ovvero SNP & Variation Suite, il quale offre un set completo di strumenti per l'elaborazione dei dati di intensità grezza, per l'identificazione delle regioni contenenti CNVs, per la visualizzazione del numero di copie e per l'esecuzione di test di associazione, di regressione lineare o logistica tra CNVs e variabili fenotipiche considerando eventuali covariate. SVS 8.7.0 impiega un algoritmo di segmentazione ottimale chiamato CNAM (Copy Number Analysis Method), utilizzando una programmazione dinamica per rilevare CNVs sia ereditate che *de novo* su una base campionaria (univariata) e multi-campione (multivariata). Alcuni algoritmi HMM presumono che i diversi stati del numero di copie debbano essere coerenti, la segmentazione ottimale, invece, delinea correttamente i confini delle CNVs.

Anche per il software SVS è necessario tenere in considerazione una serie di distorsioni del dato dovute ad effetti *batch* (piastra, macchina e variazione del sito), a *genomic waves* e/o alla stratificazione della popolazione considerata. Altre fonti di variazione includono procedure di estrazione e preparazione del campione, tipi di cellule, fluttuazioni di temperatura e persino livelli di ozono ambientale in un laboratorio.

Questi possono portare a complicazioni che vanno da segmenti scarsamente definiti a risultati falsi e non replicabili (http://goldenhelix.com/products/SNP_Variation/index.html).

Le opzioni scelte per eseguire l'analisi con SVS 8.7.0 sono state: modello univariato, un numero massimo di 100 segmenti ogni 10.000 marcatori SNPs, un numero minimo di 3 SNPs per segmento, 2.000 permutazioni a coppia con un P-value cutoff di 0,005. Anche in questo caso sono previste le correzioni delle *genomic waves* in base al contenuto GC in finestre di 1 kb ed il filtro *waviness factor* (WF) $-0,05 > WF > 0,05$.

Per le successive fasi sono stati utilizzati solo i dati provenienti dai campioni che avevano passato i filtri di qualità per entrambi i software.

Le CNVs identificate, con PennCNV e SVS 8.7.0, sono state raggruppate in CNVRs ogni qual volta si sovrapponevano all'interno dei diversi campioni, come proposto da Redon e coll. (2006) (Redon R. 2006). La procedura di raggruppamento è stata eseguita con la funzione *merge* del software BEDTools (Quinlan A.R. 2010). Infine, le CNVRs comuni ai due software sono state identificate utilizzando la funzione *intersect* di BEDTools che identifica solo le CNVRs, presenti nei due datasets, che presentano una sovrapposizione completa.

2.2. Preparazione delle librerie genomiche, sequenziamento ed analisi bioinformatiche per la validazione delle CNVs.

2.2.1 Preparazione della libreria per il sequenziamento

Per la preparazione delle librerie genomiche vengono utilizzati 50 ng di DNA per ogni individuo. In particolare, per la preparazione delle librerie, sono stati scelti 6 individui di razza Girgentana dei 36 genotipizzati con i BeadChip. Le librerie sono state preparate utilizzando il kit Nextera DNA Sample Preparation di Illumina (<https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/nextera-xt-troubleshooting-technical-note.pdf>).

In generale, il flusso di lavoro è stato condotto secondo il protocollo del produttore, come mostrato nella Figura 10.

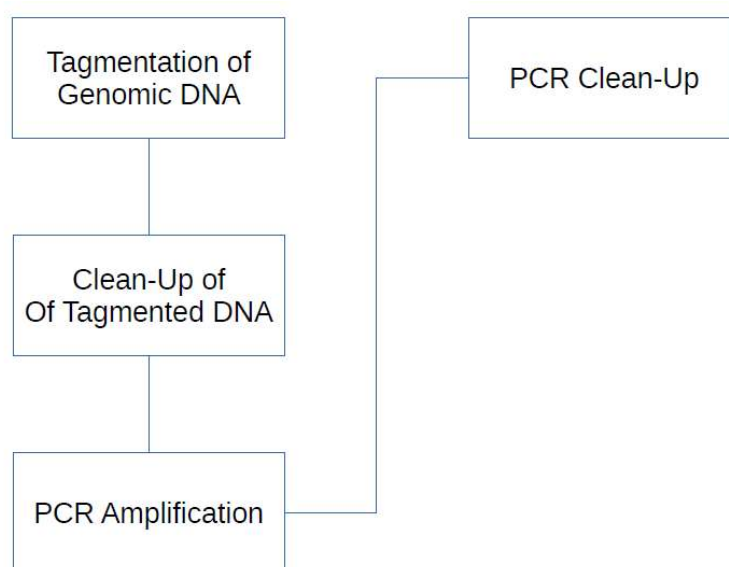


Figura 10: Flusso di lavoro per preparazione delle librerie

In breve, durante la prima fase, i campioni vengono casualmente tagmentati (ovvero taggati e frammentati) con trasposomi appositamente progettati. Il trasposome Nextera frammenta nello stesso momento il DNA genomico e aggiunge sequenze dell'adattatore alle estremità, consentendo l'amplificazione mediante PCR nei passaggi successivi. Il DNA tagmentato viene quindi amplificato tramite un programma di PCR a ciclo limitato. Vengono aggiunti marcatori di sequenza ligati su entrambe le estremità dei frammenti per identificarli e concatenarli in un secondo momento. I marcatori per ogni libreria sono stati generati con una strategia detta di doppia indicizzazione in cui si utilizzano una diversa combinazione di indici in modo tale che ogni campione sia riconosciuto (Figura 11).

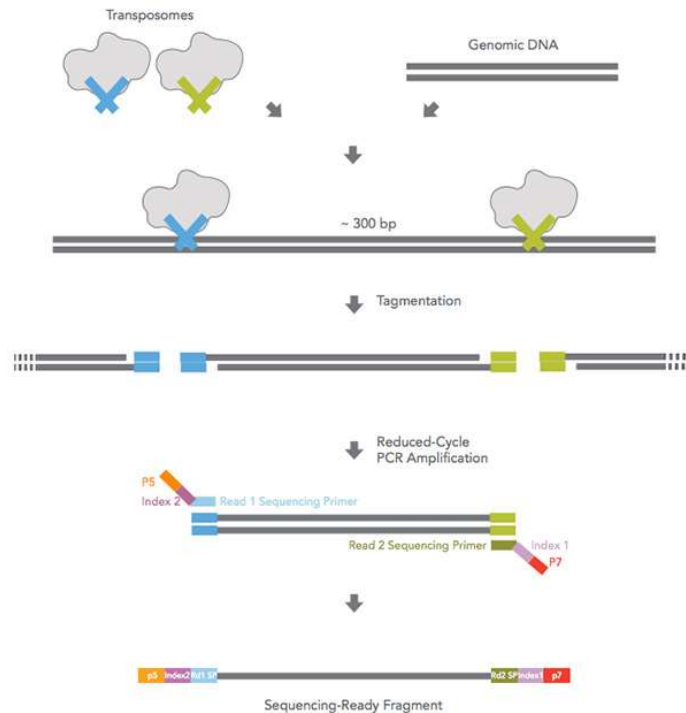


Figura 11: preparazione librerie Nextera

La fase seguente della preparazione prevede l'utilizzo di microsfere AMPure XP per purificare le librerie e rimuovere i frammenti molto brevi. Infine, le librerie vengono quantificate e diluite alle apposite concentrazioni di lavoro che variano in funzione del kit Illumina utilizzato per successiva fase di sequenziamento. Una volta che le librerie sono state preparate, sono pronte per essere sequenziate e vengono inserite all'interno di una cella di flusso che sarà poi analizzata sulla piattaforma Illumina NGS.

2.2.2 Bioinformatica in NGS

Le varie tecnologie di sequenziamento forniscono come output le sequenze nucleotidiche del campione di DNA analizzato. Delle sottostringhe di queste sequenze, di lunghezza variabile da alcune decine fino a diverse centinaia di basi, vengono combinate insieme, di solito in un file FASTQ che è un tipo di file utilizzato in biologia per memorizzare sequenze genetiche e i relativi *quality scores*, ossia il punteggio che l'algoritmo utilizzato assegna ad ogni base azotata, e che viene poi utilizzato per scegliere il *match* migliore con il genoma di riferimento. A questo punto, l'analisi bioinformatica si divide in tre passi: allineamento, che ricerca corrispondenze tra le *reads* e il genoma di riferimento, *variant calling*, che permette di individuare delle varianti strutturali (SNPs, indel, CNVs, ecc) rispetto ad un genoma di riferimento depositato in banche dati; filtraggio e annotazione, che mirano ad identificare i geni e la loro funzione biologica.

2.2.2.1 Formato FASTQ

Il NextSeq Sequencing Systems Illumina genera files di dati non elaborati nel formato BCL (*binary base call*). Come suggerisce il nome, questo viene fatto in tempo reale, cioè per ogni ciclo di sequenziamento viene aggiunta una chiamata per ogni posizione identificata sulla cella di flusso. I file BCL sono archiviati in formato binario e rappresentano l'output di dati non elaborati di una corsa di sequenziamento. Uno dei vantaggi del formato di file BCL è che ogni chiamata di base viene registrata come la macchina effettivamente chiama quella base. A differenza del formato di file FASTQ, in cui la registrazione della chiamata di base viene effettuata dopo aver letto l'intera sequenza. Però il formato BCL non è utile per nessun altro motivo se non quello di essere conveniente ed efficace per la macchina di sequenziamento pertanto questo formato di file di sequenziamento richiede la conversione in formato FASTQ per l'utilizzo con strumenti di analisi dei dati sviluppati dall'utente o da terze parti. Per poter fare questa conversione è stato usato il software **bcl2fastq** utilizzando la stringa di comando: "*bcl2fastq --input-dir/BaseCalls --samplesheet/SampleSheet.csv --runfolder-dir NOME FLOWCELL --loading-threads NUMERO DI PROCESSORI --output-dir /CARTELLA DI OUTPUT*". Per quanto riguarda il formato FASTQ questo è un formato di testo facilmente leggibile, che è stato pensato dalla Welcome Trust Sanger Institute, per associare ad una sequenza la qualità di ogni sua singola base in formato standard per l'output di strumenti di sequenziamento NGS. Il formato FASTQ è composto da quattro righe:

- 1) intestazione della sequenza, contenente l'identificatore della *read*, questa riga comincia sempre con il simbolo @;
- 2) sequenza delle basi della *reads*, di lunghezza variabile in funzione del kit di sequenziamento utilizzato, ad esempio per il sistema Illumina sono di lunghezza pari a 101 bp se abbiamo utilizzato un kit da 200 cicli;
- 3) intestazione della sequenza dei *phred values* simbolo iniziale +;
- 4) sequenza dei *phred values* corrispondenti alle basi della *read* (codificata in caratteri stampabili) (Figura 12).

```
$ head SRR094861.fastq
@SRR094861.1 HWI-EAS206_23_30H0YAAXX:5:1:709:724 length=36
ACCTCCCGCCCCNACNGCNCCTGTCACCTCCC
+SRR094861.1 HWI-EAS206_23_30H0YAAXX:5:1:709:724 length=36
5555555555555555!5!55!555555$555(.$.2.2
@SRR094861.2 HWI-EAS206_23_30H0YAAXX:5:1:1460:1251 length=36
AAACAAGCTAACATGACTAACACCTTAATTCCATC
+SRR094861.2 HWI-EAS206_23_30H0YAAXX:5:1:1460:1251 length=36
5555555555555555555555555555555555121-11
@SRR094861.3 HWI-EAS206_23_30H0YAAXX:5:1:899:1936 length=36
AACAGTCTGATTAAAAAATGGGCCAAAGAGCTTAAC
```

Figura 12: Esempio di file in formato fastq.

Ogni *phred values* q viene convertito in un carattere stampabile su file. Ogni sequenziatore NGS ha una propria funzione f che specifica la conversione dal *phred values* q :

$$c=f(q)$$

2.2.2.2 Controllo Qualità

Le moderne tecnologie di sequenziamento sono in grado di generare centinaia di milioni di sequenze in una singola corsa. Tuttavia, tali tecnologie non sono perfette, pertanto un controllo della qualità dei dati ottenuti da sequenziamento risulta essenziale al fine di garantire che i dati utilizzati per le successive analisi, non siano compromessi da sequenze di bassa qualità, sequenze contaminate o artefatti che potrebbero portare a conclusioni errate. Il controllo della qualità viene eseguito attraverso alcuni *tools* bioinformatici come il **FastQC**. In particolare il *tool* permette di controllare la qualità dei dati grezzi attraverso alcune informazioni come la lunghezza delle sequenze, i *phred quality scores*, il contenuto in GC, ecc. Questi valori vengono quindi passati ad uno strumento di visualizzazione che mostra un grafico bidimensionale (2D) delle qualità valutate. L'utente può quindi identificare visivamente dove la sequenza deve essere tagliata (trimmata) in modo da mantenere solo basi di buona qualità.

In generale vengono eliminate le basi che presentano un Q score ≤ 20 che equivale ad una probabilità del 99% di identificare le basi in modo corretto. A titolo di esempio in Figura 13 è riportato il grafico di una *read* di 101 bp di buona qualità, in quanto le singole basi hanno un Q score maggiore di 20, mentre in Figura 14 è riportata una *read* della stessa lunghezza in cui la prima base e le ultime tre hanno un Q score inferiore a 20 e saranno scartate durante la fase di *trimming*.

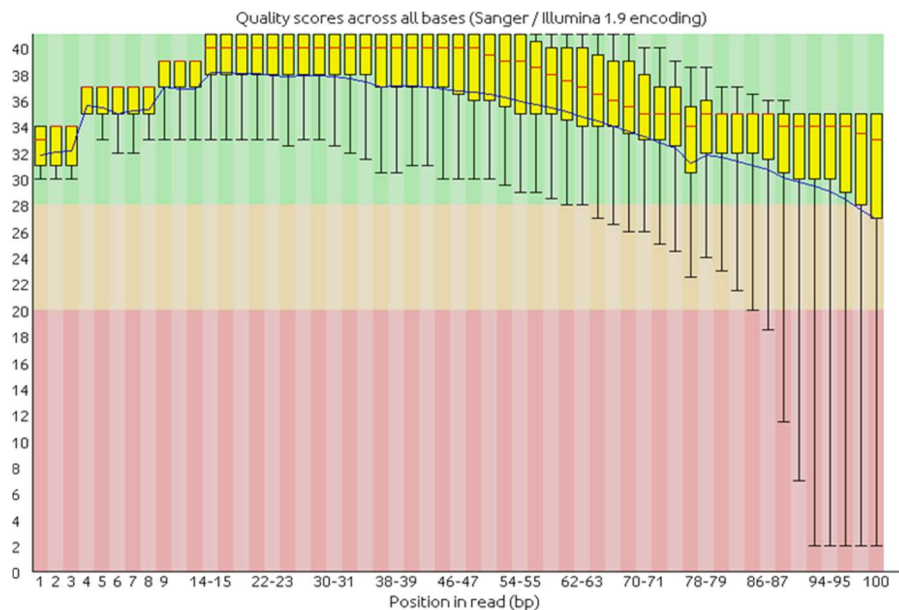


Figura 13: Distribuzione della qualità delle basi per read 19

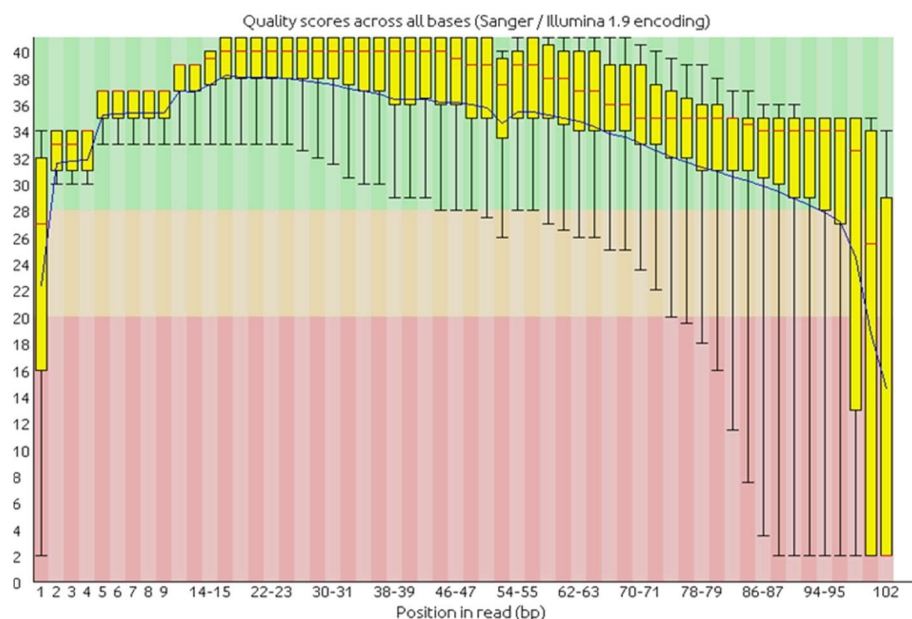


Figura 14: Distribuzione della qualità delle basi per read

2.2.2.3 Trimming ed allineamento delle reads al genoma di riferimento

La conoscenza della qualità delle *reads* consente di effettuare il *trimming* delle sequenze prima di qualsiasi processamento. Il *trimming* consiste nell'eliminare, dalle singole *reads*, quelle basi che presentano un valore di *phred score* al di sotto di una prefissata soglia ed eliminare *reads* che dopo il controllo qualità, risultano troppo corte per le successive analisi. Tra i software che è possibile utilizzare per questa operazione, uno molto utilizzato in letteratura è il **Trimmomatic**. Le *reads* che si ottengono dopo il controllo qualità e il *trimming*, devono essere quindi concatenate per ottenere una singola sequenza in formato fastq. Il *trimming* delle sequenze è stato effettuato tramite il comando: `"java -jar /Trimmomatic-0.35.jar PE-threads 8 FAZ_160_S1_L001_R1_001.fastq.gz FAZ_160_S1_L001_R2_001.fastq.gz /FAZ_160_L1_R1.paired.fastq /FAZ_160_L1_R1.unpaired.fastq /FAZ_160_L1_R2.paired.fastq /FAZ_160_L1_R2.unpaired.fastq"`.

Per quanto riguarda la fase di allineamento al genoma di riferimento, le *reads*, dopo essere state sottoposte al controllo qualità, vengono allineate al genoma di riferimento. Quest'ultimo rappresenta una sequenza di riferimento per una determinata specie che viene depositata in banche dati per poter essere utilizzata negli studi di genomica dai diversi ricercatori per i più svariati obiettivi.

Tra queste banche dati, quelle che vengono maggiormente interrogate sono Ensembl (<http://www.ensembl.org/index.html>) e UCSC, la banca dati presso la University of California Santa Cruz (<https://genome.ucsc.edu/>). La lunghezza totale del genoma di riferimento dipende dalla specie e comprende geni codificanti, geni non codificanti, pseudogeni e trascritti genici. Esso contiene inoltre varianti strutturali, come ad esempio i Copy Number Variants o inversioni, SNPs, piccole inserzioni o delezioni, e mutazioni somatiche. L'allineamento di sequenze, è una procedura bioinformatica nella quale

vengono messe a confronto ed allineate, due o più sequenze nucleotidiche di DNA. L'allineamento permette di individuare regioni identiche o simili che possono avere relazioni funzionali, strutturali o filogenetiche (evolutive), ma anche polimorfismi come SNPs e altre varianti strutturali. La fase di allineamento è tutt'altro che facile e non può essere fatta che in modo automatico attraverso un server con alte capacità computazionali. Basti pensare che il genoma bovino è costituito da circa da 3 miliardi di basi e che le *reads* ottenute da un'analisi NGS hanno una grandezza variabile da poche decine a centinaia di nucleotidi. Data la gran quantità di informazioni da gestire, il processo risulta estremamente impegnativo sia dal punto di vista informatico che dal punto di vista temporale. Esistono diversi algoritmi per allineare due sequenze tra di loro. Uno dei più utilizzati in campo zootecnico è l'algoritmo *Burrows-Wheeler transform* implementato nei *tools* bioinformatici **BWA** e **BOWTIE** che consentono di allineare *reads* corte anche laddove compaiono alcune lacune nella sequenza. In particolare, per l'allineamento delle sequenze è stata utilizzata una *pipeline* di analisi in *python* messa a punto dal gruppo di zootecnia del Dipartimento di Scienze Agrarie, Alimentari e Forestali, dell'Università degli Studi di Palermo. In particolare il software BWA è stato utilizzato per l'allineamento applicando i parametri di default ("*Map with BWA for illumina*" *tool*). Il file di formato *SAM* (*Sequence Alignment Map*) ottenuto dall'allineamento è stato successivamente convertito in *BAM* (*Binary Alignment Map*) utilizzando il *tool* "*SAM-to-BAM*" disponibile da *SAMTools*. La conversione permette di ottenere un file in linguaggio binario di dimensioni ridotte che può essere importato in maniera più veloce per le successive analisi. Dai singoli file *BAM* ottenuti per i due campioni, sono state rimosse le *reads* non mappate tramite il comando "*CleanSam*" del software Picard, in particolare tutte quelle con MAPQ = 0. Lo stesso software è stato poi utilizzato per raggruppare gli allineamenti derivati da una coppia di *reads* tramite il comando "*SortSam*" con l'opzione "*queryname*" e successivamente per ordinarli per coordinate cromosomiche e posizione tramite l'opzione "*coordinate*". L'applicazione "*MarkDuplicates*" è stata infine utilizzata per contrassegnare e rimuovere le *reads* mappate in regioni multiple (duplicate). Tra le statistiche che vengono applicate dopo la fase di allineamento, molto importante è la stima del *coverage* (copertura), *c*, dato dal rapporto tra il prodotto del numero totale delle *reads* (*n*) e la lunghezza media delle *reads* (*l*) e la lunghezza totale della sequenza assemblata (*L*) (Figura 15): $c = n \cdot l / L$; maggiore è il *coverage* di una sequenza più robusti sono i risultati che si otterranno nella fase di analisi successiva.

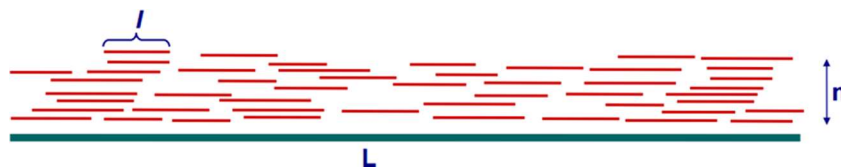


Figura 15: Rappresentazione grafica del coverage di una sequenza.

Il file BAM risultante è pronto per essere utilizzato nelle fasi successive delle analisi bioinformatiche.

2.2.2.4 CONTROL-FREEC

Control-FREEC (Control Copy Copy Number e allelic content caller) è uno strumento per il rilevamento delle CNVs e degli squilibri allelici (incluso LOH) utilizzando dati di sequenziamento originariamente sviluppati dal Laboratorio di Bioinformatica dell'Institut Curie (Parigi). Control-FREEC calcola, normalizza, segmenta automaticamente il numero di copie e i profili di B Allele Frequency (BAF), quindi chiama il numero di copie e LOH. Il software è in grado di rilevare automaticamente le alterazioni del numero di copie (*Copy Number Alteration*, CNA) anche senza utilizzare un set di dati di controllo (Boeva V. 2011).

Control-FREEC prende come file di input sequenze allineate (es. SAM o BAM), quindi costruisce e normalizza il profilo del numero di copie, costruisce il profilo di frequenza dell'allele B (BAF), segmenta entrambi i profili, attribuisce lo stato genotipico a ciascun segmento utilizzando sia il numero di copie che le informazioni di frequenza allelica, quindi annota le alterazioni genomiche. Se è disponibile un campione di controllo Control-FREEC rileva varianti somatiche da quelle germinali. Control-FREEC definisce con precisione lo stato del genotipo anche quando non è disponibile alcun esperimento di controllo o se il genoma è poliploide. Corregge quelli che sono i pregiudizi relativi al contenuto GC e alla mappabilità.

Le opzioni utilizzate per il file di configurazione per eseguire l'analisi con Control-FREEC sono state: il *chrFiles* in cui viene specificato il percorso del file contenente il numero di cromosomi (da 1 a 29 nel caso del genoma caprino) e la relativa lunghezza in paia di basi; *ploidy* è stato inserito un valore pari a 2; *chrLenFile* in cui è specificato il percorso della cartella contenente i singoli cromosomi in formato fasta; *mateFile* dove viene indicato il percorso del file delle sequenze allineate ossia il file BAM realizzato in precedenza, infine alla voce *window* come da manuale è stato dato un valore pari a 50.000 bp. Lanciata l'analisi tra i file di output che si ottengono vi è l'*info.txt* il quale contiene informazioni sulla corsa, il file *CNV* che presenta le coordinate delle CNVs predette, con informazioni riguardanti il cromosoma, con la posizione di inizio e fine, con il tipo di alterazione, con il numero di copie predetto e con informazioni riguardanti il genotipo; il file *ratio.txt* il quale è un file con i rapporti e le alterazioni del numero di copie previste per ogni finestra; l'output *sample.cnp* il quale mostra i files con profili di CNVs grezzi; il *GC_profile.cnp* il quale non è altro che un file con il profilo del contenuto di GC. Infine, è stato utilizzato uno script di R modificato per il genoma caprino per la visualizzazione delle CNVs lungo il genoma utilizzando, il *makeGraph.R*.

Questo strumento è stato utilizzato per validare le CNVRs *consensus* ovvero quelle presenti nei due datasets di PennCNV e SVS con una sovrapposizione del 99%.

2.2.3 Annotazione di geni all'interno delle CNVRs validate

L'annotazione genica delle CNVRs *consensus*, che sono state validate con il sequenziamento, è stata effettuata basandosi sulla sequenza di riferimento del genoma caprino GCF_001704415.1 anche nota come *assembly* ARS1, attraverso l'utilizzo del browser Genome Data Viewer del National Center for Biotechnology Information Database (https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?context=genome&acc=GCF_001704415.1).

Capitolo III: Risultati e discussione

1 Risultati CNVs provenienti da dati di genotyping

1.1 Risultati PennCNV e SVS

Una volta lanciata l'analisi con i due software si sono ottenuti i seguenti risultati. La segmentazione univariata di SVS ha individuato 831 CNVs, delle quali 732 erano delezioni (*Loss*) e 99 duplicazioni (*Gain*). Il modello HMM del software PennCNV ha identificato 230 CNVs delle quali 55 *Loss* e 175 *Gain*. Il primo filtro applicato alla chiamata delle CNVs è stata l'eliminazione di tutti gli animali (e quindi delle relative CNVs) che non erano contemporaneamente presenti in entrambi i dataset. Al termine di quest'ultimo filtro sono rimaste 804 CNVs per le analisi condotte con SVS, delle quali 712 *Loss* e 92 *Gain* mentre le CNVs ottenute col software PennCNV sono 192 delle quali 39 *Loss* e 153 *Gain*. In Tabella 2 sono riportate le statistiche descrittive delle CNVs ottenute con i due software.

Tipo	CNVs	Lunghezza media	Lunghezza mediana	Size range	Min CNVs per campione	Max CNVs per campione
SVS						
Loss	712	241.746	106.510	9.227-3.579.727	14	34
Gain	92	330.032	5.212.315	6.986-10.297.329	1	11
Totale	804	251.849	106.510	6.986-10.297.329	18	34
PennCNV						
Loss	39	179.879	122.999	46.687-957.268	1	7
Gain	153	233.153	122.999	14.570-1.864.334	2	11
Totale	192	222.332	237.771	14.570-1.864.334	3	12

Tabella 2: Caratteristiche e dimensioni (bp) delle CNVs evidenziate con i software SVS e PennCNV

Dai risultati della Tabella 2, si può notare che PennCNV identifica un numero maggiore di *Gain* rispetto al totale, al contrario del software SVS che identifica più *Loss*. Il valore medio delle CNVs per campione è di 6,4 per PennCNV, con valori medi di 1,95 e 5,1 rispettivamente per le CNVs *Loss* e *Gain*. Il valore medio delle CNVs per campione è di 26,8 per SVS, con valori medi di 23,73 e 4,38 rispettivamente per le CNVs *Loss* e *Gain*.

Il rapporto di delezione (*deletion ratios*) calcolato come il numero totale delle CNVs *Loss* diviso il numero totale delle CNVs è del 20,31% e 88,56% rispettivamente per PennCNV e per SVS.

Si può notare che le CNVs individuate con il software SVS sono generalmente più lunghe.

1.2 CNVRs e definizione di consensus CNVRs

Le CNVs identificate, con PennCNV e SVS 8.7.0, sono state raggruppate in CNVRs ogni qual volta si sovrapponevano all'interno dei diversi campioni, come proposto da Redon e coll. (2006) (Redon R. 2006).

Le CNVRs trovate con il software SVS sono 361 con 312 CNVRs *Loss*, 43 *Gain* e 6 *Complex*, ovvero sono sia *Loss* che *Gain*. Le CNVRs presentavano una lunghezza media pari a circa 341.5 kb, infine il range totale di lunghezza andava da un minimo di 9.834 bp a un massimo di circa 10.297 kb (vedi Tabella 3). La maggior parte delle CNVRs identificate erano localizzate sul cromosoma 1, in particolare il 6,09%, mentre i cromosomi caprini con meno CNVRs erano il 10, 23 e 28 con un numero di CNVRs pari a 7 (1,94%) (Figura 16).

Le CNVs identificate con PennCNV sono state raggruppate in 59 CNVRs, delle quali 28 CNVRs erano *Loss*, 29 *Gain* e 2 *Complex*. Il range di lunghezza delle CNVRs andava da un minimo 48.775 bp a un massimo di circa 1.864 kb. La maggior parte delle CNVRs erano localizzate sui cromosomi 12 e 15 (l'8,50%), mentre i cromosomi 10, 17, 19, 20, 24 e 26 presentavano un numero di CNVR pari a 1 (1,69%) (Figura 16).

L'analisi delle CNVRs *consensus*, ovvero delle CNVRs comuni ai due software, ha identificato 15 CNVRs. Tra queste 5 CNVRs erano *Gain*, 9 *Loss* e una *Complex*.

Il range della lunghezza delle CNVRs *consensus* va da 75.025 bp a 1.253.262 bp con una lunghezza media totale pari a circa 296.647 bp. La distribuzione sui cromosomi delle CNVRs identificate da PennCNV, SVS e dalle CNVRs *consensus* è riportata in Figura 16.

Tipo	CNVRs	Lunghezza media	Lunghezza mediana	Size range
SVS				
Loss	312	249.988,6	127.088,5	43.854-3.291.743
Gain	43	639.898,4	248.800	9.834-3.408.146
Complex	6	296.0748,2	1.471.891,5	36.925-1.0297.329
Totale	361	341.486,3241	131.461	9.834-10.297.329
PennCNV				
Loss	28	163.639,6	115.803	77.671-688.365
Gain	29	274.964,2	126.898	48.775-1.864.334
Complex	2	260.989,5	260.989,5	59.089-462.890
Totale	59	221.658,5	125.366	48.775-1.864.334
Consensus				
Loss	9	173.840,8	104.418,5	78.908-688.365
Gain	5	547.176,4	248.800	75025-1253262
Complex	1	149.264	149.264	149264
Totale	15	296.647,5	111.757	75025-1253262

Tabella 3: Statistiche descrittive delle CNVRs identificate dai software SVS, PennCNV e dalla consensus analisi

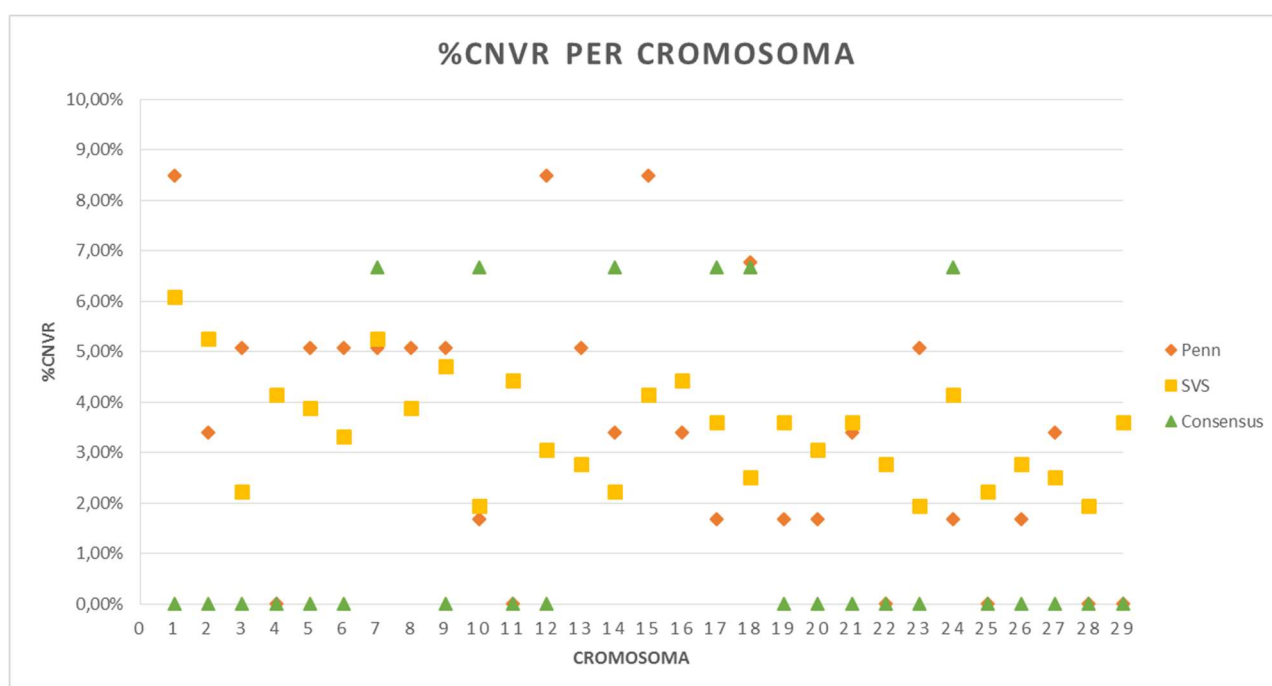


Figura 16: Percentuale di CNVRs per i cromosomi caprini ottenuti con i software SVS, PennCNV e con la consensus analisi

Inoltre, date le differenze nelle proporzioni delle lunghezze delle CNVRs individuate attraverso la *consensus* analisi e attraverso i due software si è operata una comparazione tra le lunghezze. Le CNVRs di SVS, PennCNV e della *consensus* analisi sono state suddivise in 5 classi di lunghezza, per ogni stato. Queste cinque classi come si può notare in Figura 17 erano così suddivise:

- 1-5 Kb;
- 5-10 Kb;
- 10-100 Kb;
- 100-1000 Kb
- > 1000 Kb

La maggior parte delle CNVRs identificate in questo studio avevano una lunghezza compresa tra le 100 e 1000 Kb, principalmente per le CNVRs identificate con SVS e con PennCNV per quanto riguarda i due stati *Gain* e *Loss*, mentre per quanto riguarda le *Complex* si può notare che la maggior parte delle CNVRs erano distribuite non solo tra 100 e 1000 Kb ma anche sopra 1000 Kb, principalmente per quelle identificate con il software SVS (66,70%). Per quanto riguarda le CNVRs identificate con la *consensus* analisi presentavano una distribuzione differente rispetto a quelle identificate con i due software. Per quanto riguarda lo stato *Gain*, le CNVRs erano maggiormente distribuite tra i 10-100 Kb e > 1000 Kb, per lo stato *Complex* queste avevano totalmente una lunghezza che andava tra le 100 e le 1000 Kb, mentre per lo stato *Loss* erano distribuite tra 10-100Kb e 100-1000 Kb.

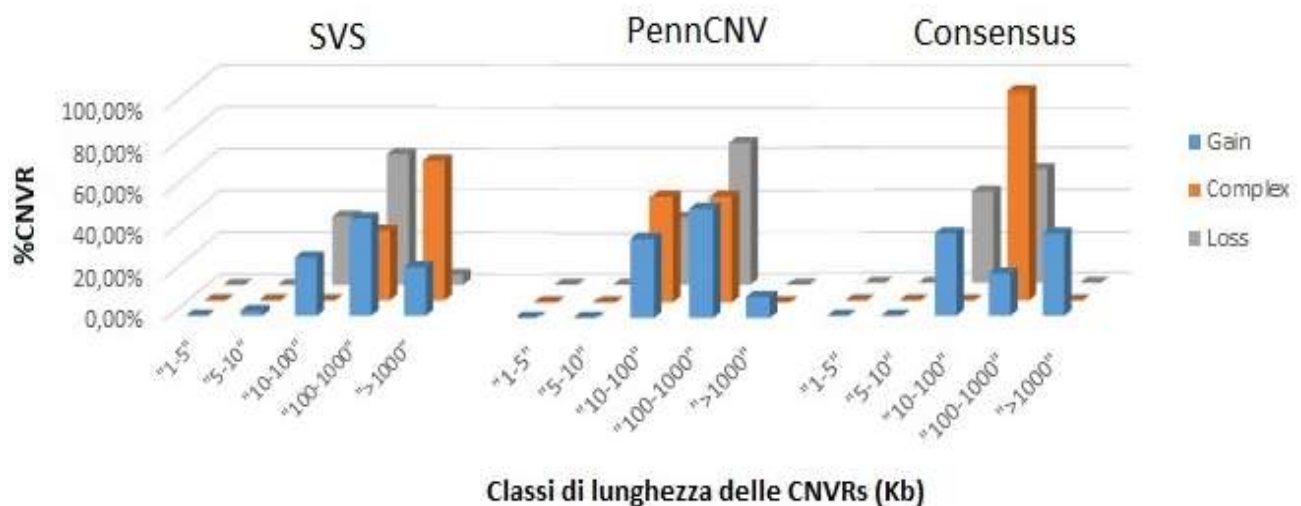


Figura 17: Distribuzione delle lunghezze delle CNVRs identificate con SVS, PennCNV e con la *consensus* analisi.

In Figura 18 vengono riportate la distribuzione del numero delle CNVRs che rientrano in una specifica classe di lunghezza. Come si può notare per entrambi i software, PennCNV e SVS, le CNVRs (che poi definiscono le CNVRs *consensus*) più rappresentate sono quelle delle regioni *Loss* le quali hanno una lunghezza che rientra nel range che va da 100 Kb a 1000 Kb.

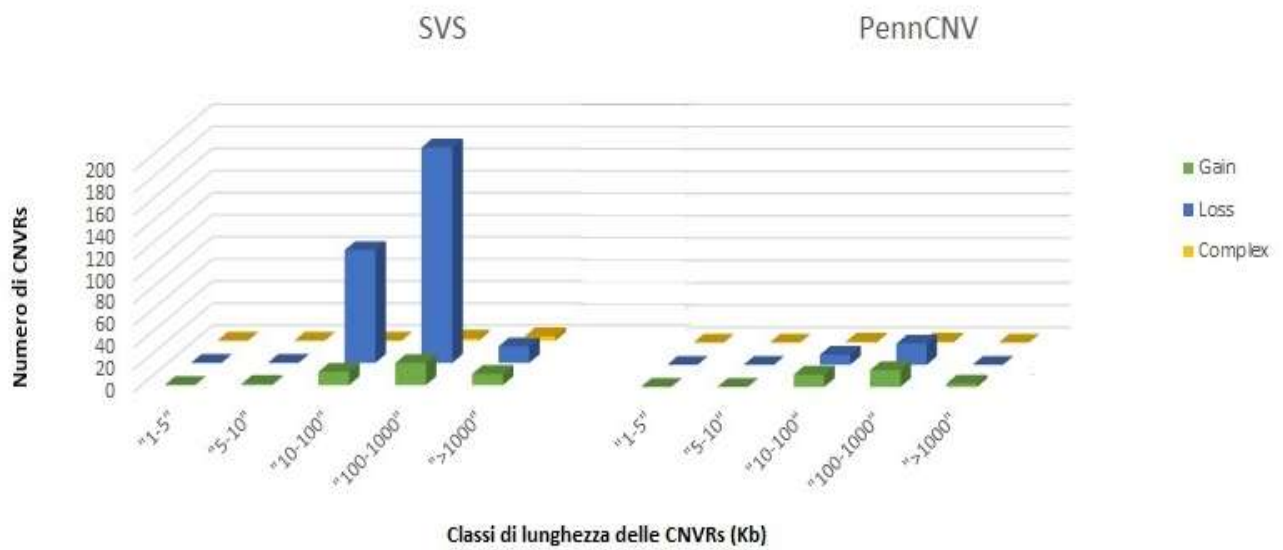


Figura 18: Distribuzione delle CNVRs, identificate con SVS e PennCNV, basata sulle classi di lunghezza delle CNVRs.

Nelle Figure 19, 20 e 21 vengono mostrate le caratteristiche calcolate sui campioni per ogni stato delle CNVRs, in accordo anche alle classi di lunghezza indicate precedentemente. I campioni sono stati suddivisi in 5 classi:

- 1
- 2-5
- 6-15
- 15-20
- >20

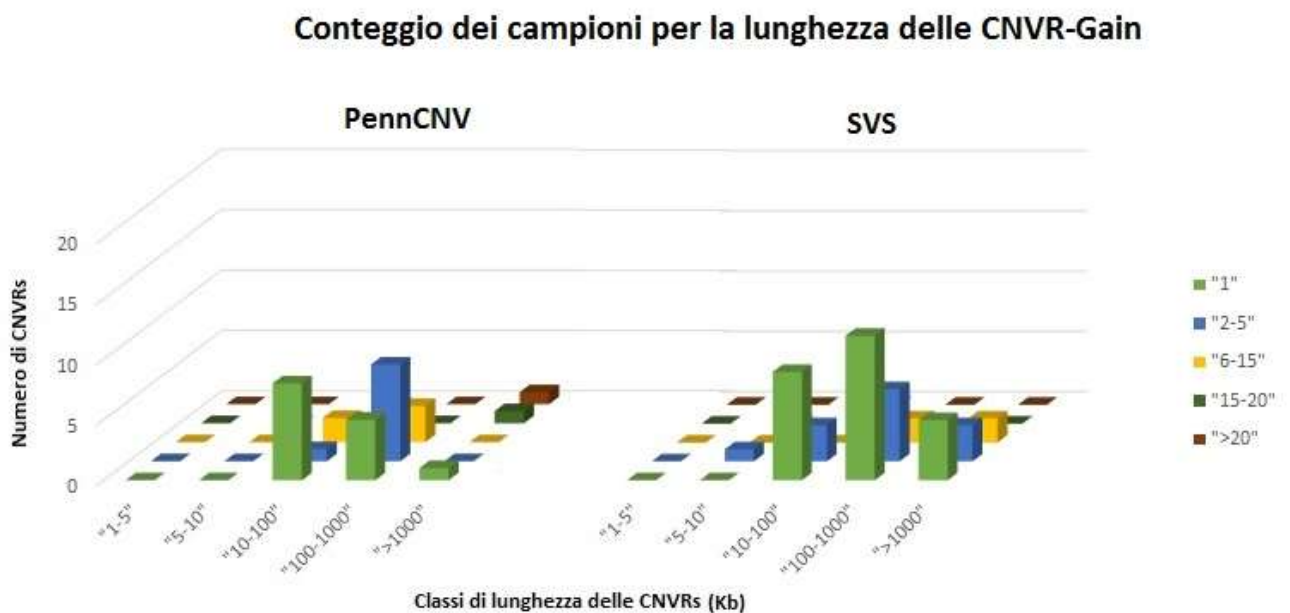


Figura 19: Conteggio dei campioni per la lunghezza delle CNVR in base allo stato *Gain*

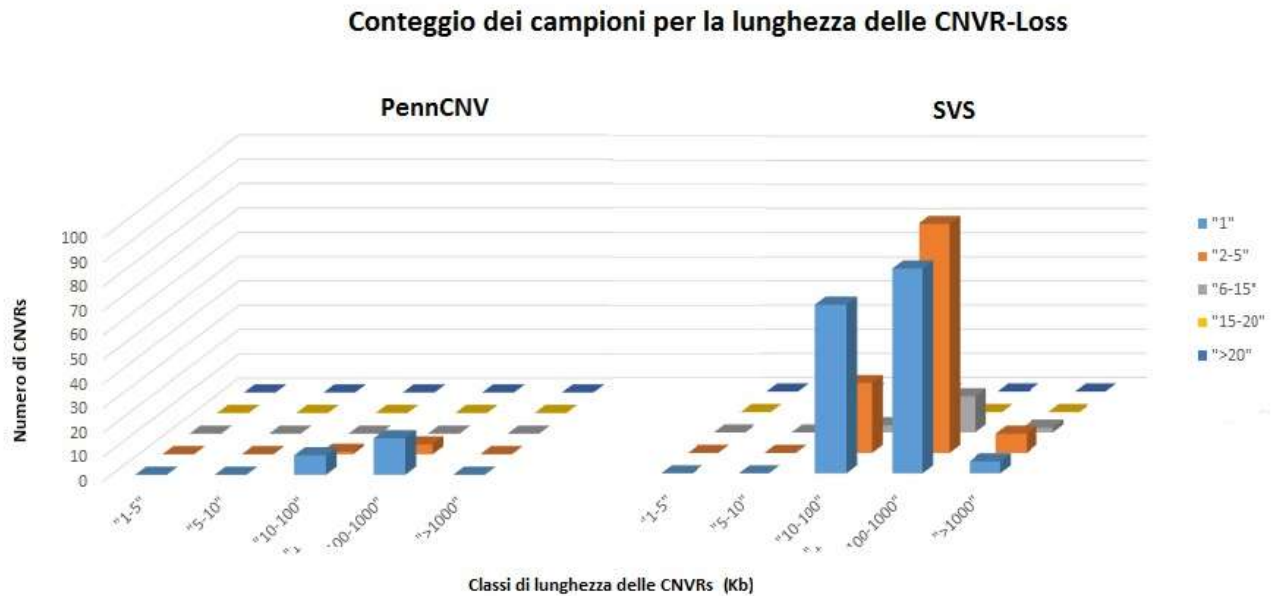


Figura 20: Conteggio dei campioni per la lunghezza delle CNVR in base allo stato *Loss*

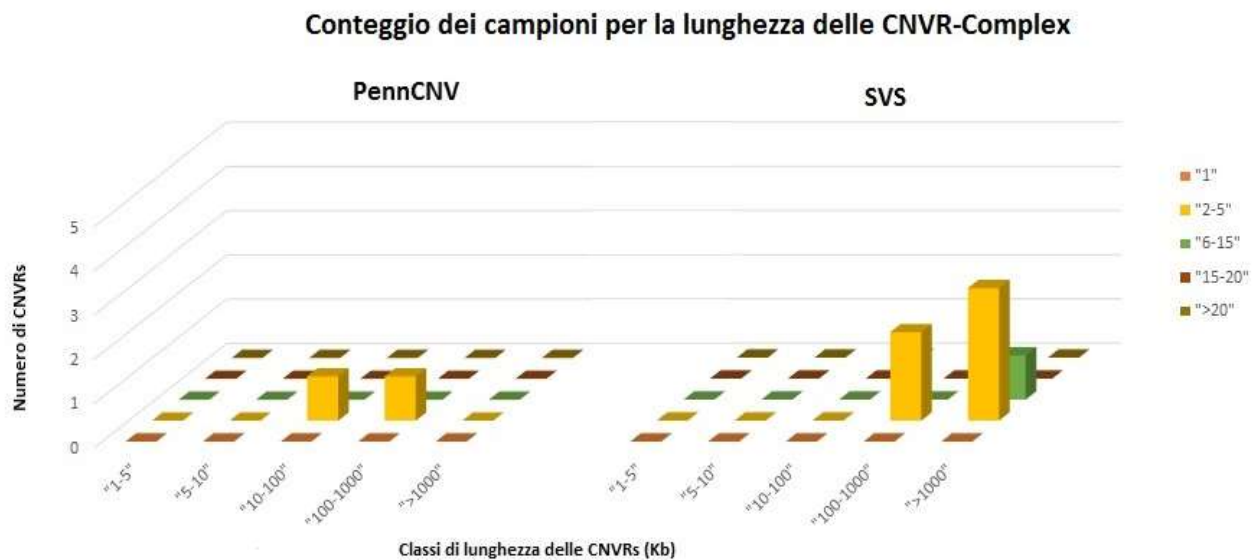


Figura 21: Conteggio dei campioni per la lunghezza delle CNVR in base allo stato *Complex*

Come si può notare, si hanno sempre più campioni lungo le classi di lunghezza 10-100 Kb e 100-1000 Kb, sia per quanto riguarda PennCNV sia per SVS. Ovviamente come si evince dai grafici si ha un numero maggiore di CNVRs allo stato *Loss* presenti maggiormente nella classe di lunghezza 100-1000 Kb. Inoltre per entrambi i software la classe di campioni principale è quella pari a 1 che è appunto la più frequente tra le classi.

Per quanto riguarda PennCNV, la classe 1 si distribuisce maggiormente tra le classi di lunghezza 10-100 e 100-1000 Kb, sia per lo stato *Gain* che per lo stato *Loss*. Inoltre per quanto riguarda la classe 2-5 questa si distribuisce maggiormente nella classe di lunghezza 100-1000 Kb, sia per lo stato *Gain* che per lo stato *Loss*.

Per quanto riguarda SVS, si può notare come la classe 1 si distribuisca diversamente rispetto a quanto avviene in PennCNV. Per lo stato *Gain* la classe 1 si ritrova dalla classe di lunghezza 10-100 Kb fino alla classe di lunghezza > 1000 Kb con un picco massimo nella classe di lunghezza 100-1000 Kb; per quanto riguarda invece quello che accade nello stato *Loss*, la classe 1 si distribuisce tra le classi 10-100, 100-1000 e > 1000 Kb ma assume valori simili tra le classi di lunghezza 10-100 e 100-1000 Kb. Per la classe di campioni 2-5, sia per lo stato *Gain* che per lo stato *Loss* questa è maggiormente presente lungo la classe di lunghezza 100-1000 Kb.

Rispetto ai due stati *Gain* e *Loss* diversa è la situazione dello stato *Complex*, in PennCNV si nota che la classe di campioni principalmente presente, dato che le altre sono nulle, è quella 2-5 la quale si distribuisce in ugual misura tra le classi di lunghezza 10-100 e 100-1000 Kb; in SVS invece si nota, così come in PennCNV, che le classi 1, 15-20 e > 20 sono assenti mentre la classe 2-5 si distribuisce tra le classi di lunghezza 100-1000 e >1000 Kb con valori simili, inoltre è presente anche la classe 6-15 che è presente principalmente nella classe di lunghezza >1000 kb.

Infine, come si può evincere dalla Tabella 4 basata sui dati della *consensus* analisi, vengono riportate le classi dei campioni, le classi di lunghezza e il numero delle CNVRs per ogni classe sia di lunghezza che di campione e come si può notare vi è una distribuzione simile tra i due software soprattutto per quanto riguarda lo stato *Loss* in cui si nota una distribuzione simile della classe 1 lungo le classi di lunghezza 10-100 e 100-1000 Kb.

PennCNV							SVS				
		"1"	"2-5"	"6-15"	"15-20"	">20"	"1"	"2-5"	"6-15"	"15-20"	">20"
GAIN	"1-5"	0	0	0	0	0	0	0	0	0	0
	"5-10"	0	0	0	0	0	0	0	0	0	0
	"10-100"	1	0	1	0	0	0	2	0	0	0
	"100-1000"	0	1	0	0	0	1	0	0	0	0
	">1000"	0	0	0	1	1	0	1	1	0	0
LOSS		"1"	"2-5"	"6-15"	"15-20"	">20"	"1"	"2-5"	"6-15"	"15-20"	">20"
	"1-5"	0	0	0	0	0	0	0	0	0	0
	"5-10"	0	0	0	0	0	0	0	0	0	0
	"10-100"	3	1	0	0	0	4	0	0	0	0
	"100-1000"	4	1	0	0	0	2	2	1	0	0
COMPLEX	">1000"	0	0	0	0	0	0	0	0	0	0
		"1"	"2-5"	"6-15"	"15-20"	">20"	"1"	"2-5"	"6-15"	"15-20"	">20"
	"1-5"	0	0	0	0	0	0	0	0	0	0
	"5-10"	0	0	0	0	0	0	0	0	0	0
	"10-100"	0	0	0	0	0	0	0	0	0	0
	"100-1000"	1	0	0	0	0	0	1	0	0	0
	">1000"	0	0	0	0	0	0	0	0	0	0

Tabella 4: Classi dei campioni, classi di lunghezza e numero delle CNVRs per ogni classe sia di lunghezza che di campione ottenuti dalla consensus analisi.

2 Dati di sequenziamento

2.1 Processamento delle reads e Quality Check

Il sequenziamento dell'intero genoma è stato eseguito con la piattaforma Illumina, dopo aver preparato la libreria con il kit Nextera (Illumina s.d.). Ogni libreria è stata sequenziata su 4 singole *lane* di una *flowcell*. Un totale di 135.540.363 di *clusters* sono stati generati, mentre quelli che hanno passato il filtro sono stati 124.130.026, con una dimensione di 37.487 Mb ed una dimensione media di 9.372 Mb.

Lane	PF Clusters	% of the lane	% Perfect barcode	% One mismatch barcode	Yield (Mbases)	% PF Clusters	% >= Q30 bases	Mean Quality Score
1	33.173.301	100.00	97.80	2.20	10.018	92.24	84.32	32.74
2	29.720.738	100.00	96.86	3.14	8.976	89.42	81.64	32.25
3	30.742.550	100.00	97.60	2.40	9.284	91.87	81.06	32.15
4	30.493.437	100.00	98.18	1.82	9.209	92.76	82.82	32.47

Tabella 5: Dati flowcell.

Per quanto riguarda il primo campione sequenziato, denominato FAZ_160, 52.678.531 *clusters* hanno passato il filtro con una dimensione totale di 15.909 Mb, con una media di 5.295 Mb; mentre per quanto riguarda il secondo campione denominato FAZ_198, 66.758.815 *clusters* hanno passato il filtro con una dimensione totale di 20.161 Mb e una dimensione media di 3.905 Mb. Inoltre, per quanto riguarda la qualità di sequenziamento dei singoli campioni, si ritrova un valore di “*Mean Quality Score*” uguale al 32,55 per FAZ_160 e 32,43 per FAZ_198. La prima analisi ha riguardato la conversione dei files *.bcl* generati dal sequenziamento, in file *fastqc* tramite l'utilizzo del *tool* informatico fornito da Illumina: **bcl2fastq.pl**.

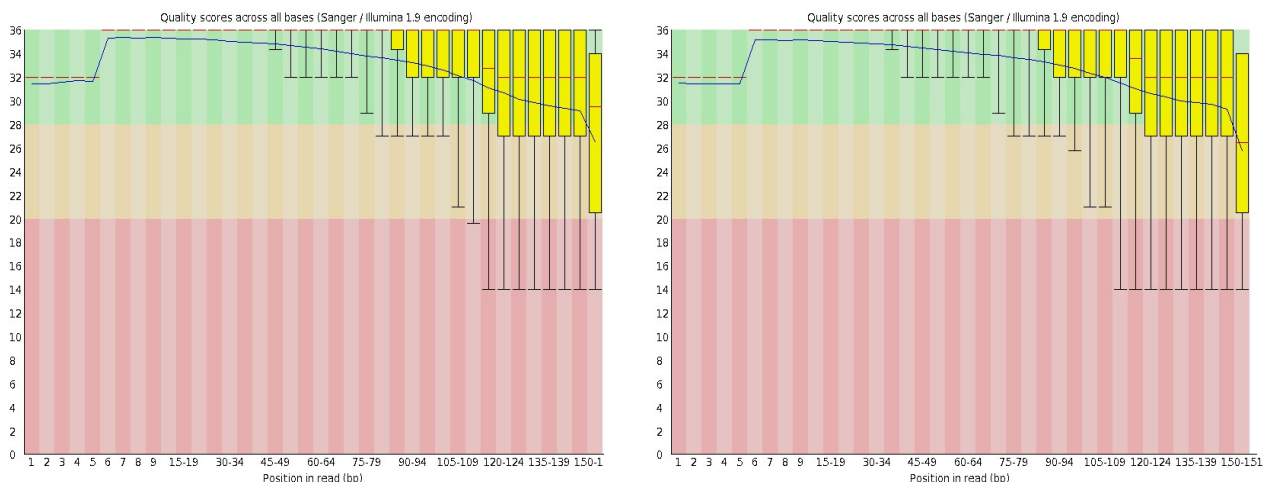


Figura 22: qualità media di ogni posizione nelle reads per il campione FAZ_160 forward FAZ_160 reverse

A titolo di esempio nella Figura 22 sono riportate le sequenze *forward* e *reverse* prima di filtrarle per la qualità, mentre nella Figura 23 si può notare come la qualità media per ogni singola base sia migliorata sia per il *forward* che per il *reverse*.

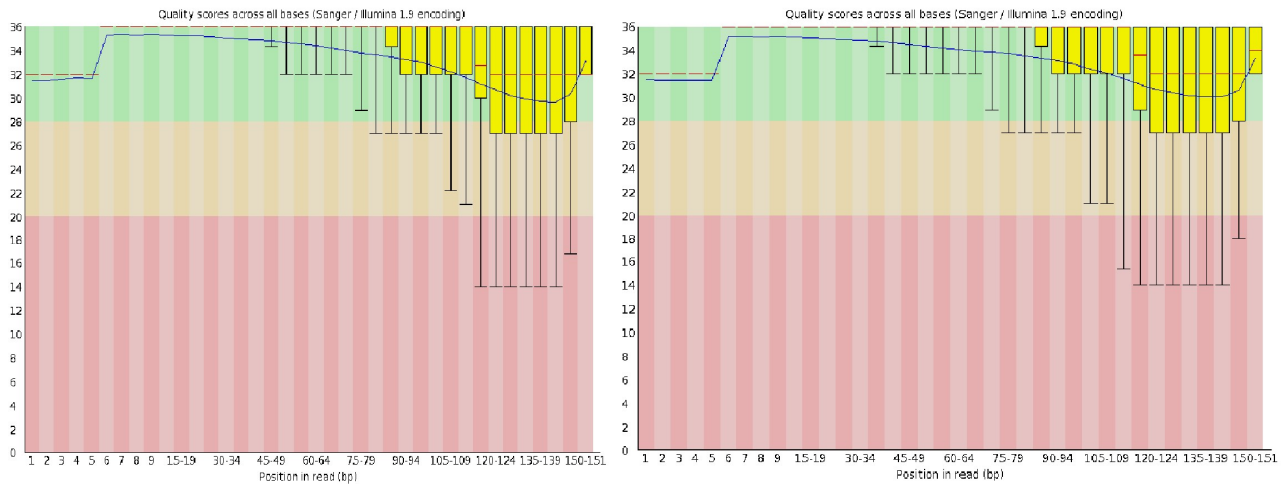


Figura 23: qualità media di ogni posizione nelle reads, dopo trimmaggio, per il campione FAZ_160 forward e FAZ_160 reverse.

2.2 Mappaggio e allineamento

Le sequenze filtrate per la qualità sono state utilizzate come input per l'allineamento col genoma caprino di riferimento e per le analisi successive. Il comando “*flagstat*” del software **SAMTools** è stato utilizzato per ottenere delle statistiche sulle *reads* mappate al genoma caprino (Tabella 6).

	FAZ_160	FAZ_198
QC-passed reads + QC-failed reads	104.566.247 + 0 in total	99.956.625 + 0 in total
mapped	104.212.712 + 0 (99.66%:-nan%)	99.138.671 + 0 (99.18%:-nan%)
paired in sequencing	104.158.229 + 0	99.446.625 + 0
read1	52.076.736 + 0	49.720.068 + 0
read2	52.081.493 + 0	49.726.557 + 0
properly paired	103.031.824 + 0 (98.92%:-nan%)	97.483.040 + 0 (98.03%:-nan%)
with itself and mate mapped	103.723.708 + 0	98.529.846 + 0
singletons	80.986 + 0 (0.08%:-nan%)	98.825 + 0 (0.10%:-nan%)

Tabella 6: statistiche delle sequenze dopo allineamento.

Tra le *reads* che hanno passato il *Quality Control*, 104.212.712 di *reads* sono state mappate per FAZ_160 (99,66%) e 99.138.671 per FAZ_198 (99,18%), tra queste quelle accoppiate durante il sequenziamento sono 104.158.229 e 99.446.625, rispettivamente per i due campioni. Questo dato dell'accoppiamento è dato dalla somma della *read1* e della *read2*. Di queste il 98,92% e il 98,03% sono risultate *properly paired*, ossia entrambi gli accoppiamenti di una coppia di *reads* mappano sullo stesso cromosoma ed orientati in maniera corretta. I *Singletons* in cui solo un *tag* è mappato correttamente è uguale allo 0,08% e 0,10% rispettivamente per FAZ_160 e FAZ_198. Infine, dai file *flagstat* si può dire che entrambe le mappature

sono abbastanza simili, per quanto riguarda la percentuale di coppie e *singleton* correttamente mappati. La profondità del sequenziamento indica quante volte in media ogni base è stata letta durante il sequenziamento. Nel presente lavoro, la profondità media di sequenziamento è stata stimata con il comando *depth* del software *samtools* (“*samtools depth aln.sorted.bam > depth.txt*”) e con il seguente script: “*awk '{sum+=\$3} END { print "Average = ",sum/NR}' depth.txt*” che ha restituito un valore di profondità di 28,4 e di 27,2 per i campioni FAZ_160 e FAZ_190, rispettivamente. Come si evince dalla Tabella 6 la copertura media sul genoma di riferimento (*mapped reads*) è stata di 99,66% e 99,18% per i campioni FAZ_160 e FAZ_198 rispettivamente.

2.3 Chiamata e Validazione delle CNVs

Una volta allineate le *reads* dei due campioni al genoma di riferimento si è proceduti, attraverso il software Control-FREEC v.11.5, alla identificazione delle CNVs presenti all’intero dei nostri due campioni FAZ_160 e FAZ_198. Dal file di output “*info.txt*” relativo ai due campioni è possibile vedere che l’algoritmo utilizzato dal software converge e sono state utilizzate 90.325.396 e 84.255.386 coppie di *reads* per FAZ_160 e FAZ_198, rispettivamente. Come accennato in precedenza sono state utilizzate finestre genomiche di 50.000 bp ed il contenuto GC per normalizzare le CNVs identificate in assenza di un campione controllo. Al termine dell’analisi nel campione FAZ_160 sono state identificate 284 CNVs con un minimo di 2 nel cromosoma 18 ad un massimo di 19 nel cromosoma 8. Di queste CNVs, 123 hanno uno status di *Gain* e 161 di *Loss*. Nel secondo campione FAZ_198 sono state identificate 163 CNVs con un minimo di 1 sul cromosoma 28 ed un massimo di 13 sul cromosoma 3. Di queste 163 CNVs, 104 hanno uno status di *Gain* mentre 59 hanno uno status di *Loss* (Tabella 7). Confrontando i risultati ottenuti con CONTROL-FREEC e quelli ottenuti dai due software SVS e PENN-CNV, possiamo osservare che i risultati ottenuti dall’analisi delle sequenze dei due campioni di capra, si avvicinano a quelli di PENN-CNV, mentre SVS tende a sovrastimare le CNVs identificate.

Campione	# CNV	min/chr	max/chr	Gain	Loss
FAZ_160	284	2/18	19/8	123	161
FAZ_198	163	1/28	13/3	104	59

Tabella 7: Numero di CNVs identificati e status nei due campioni

Nelle Figure 24 e 25 sono riportate le CNVs identificate per cromosoma nei due campioni di razza Girgentana considerati per la validazione. Sull’asse delle ascisse è riportata la posizione in bp all’interno

del cromosoma e sull'asse delle ordinate il profilo normalizzato delle CNVs che sono disegnate in rosso in caso di status *Gain* ed in blu per lo status *Loss*.

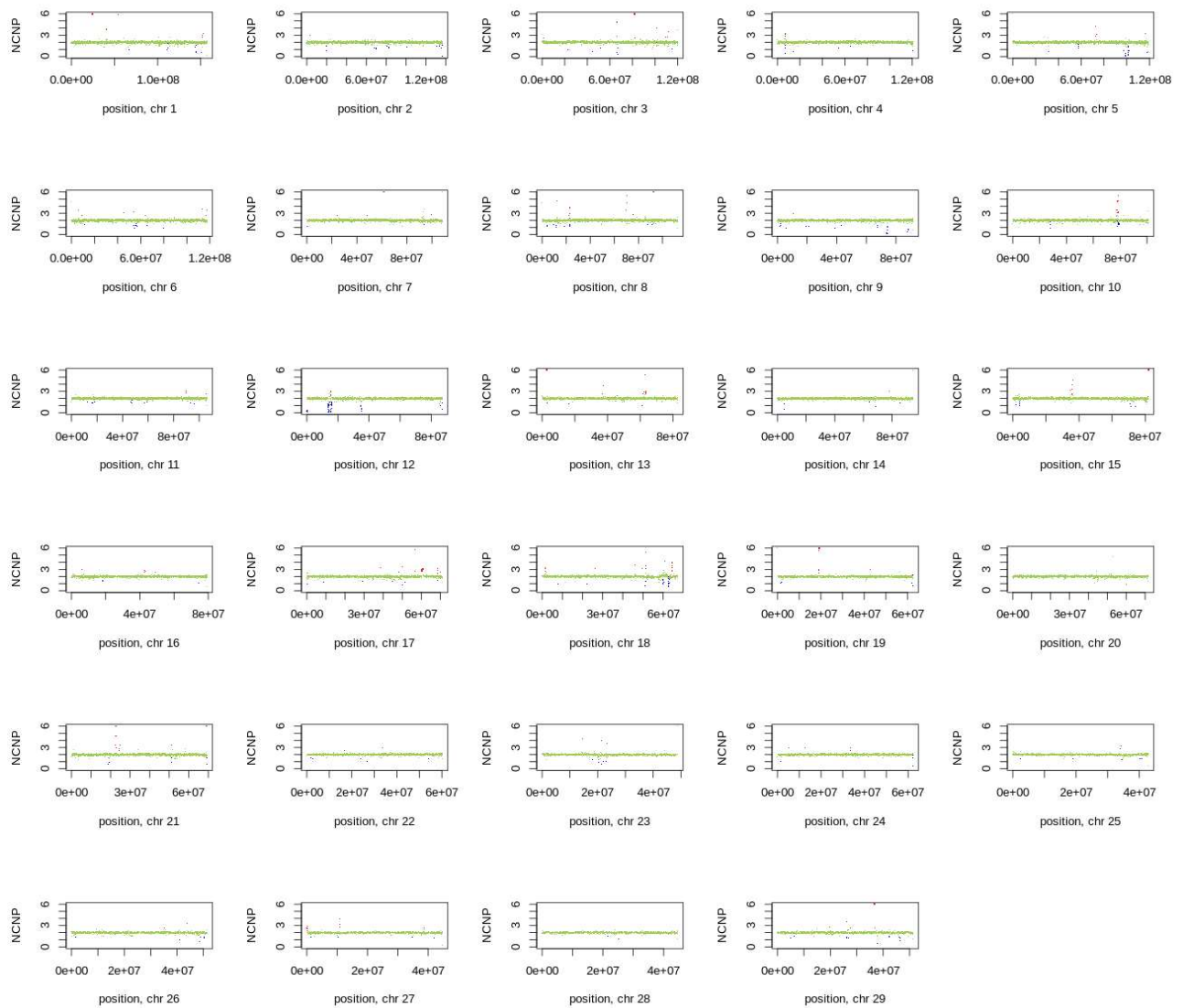


Figura 24: CNVs identificati nel genoma del campione FAZ_160

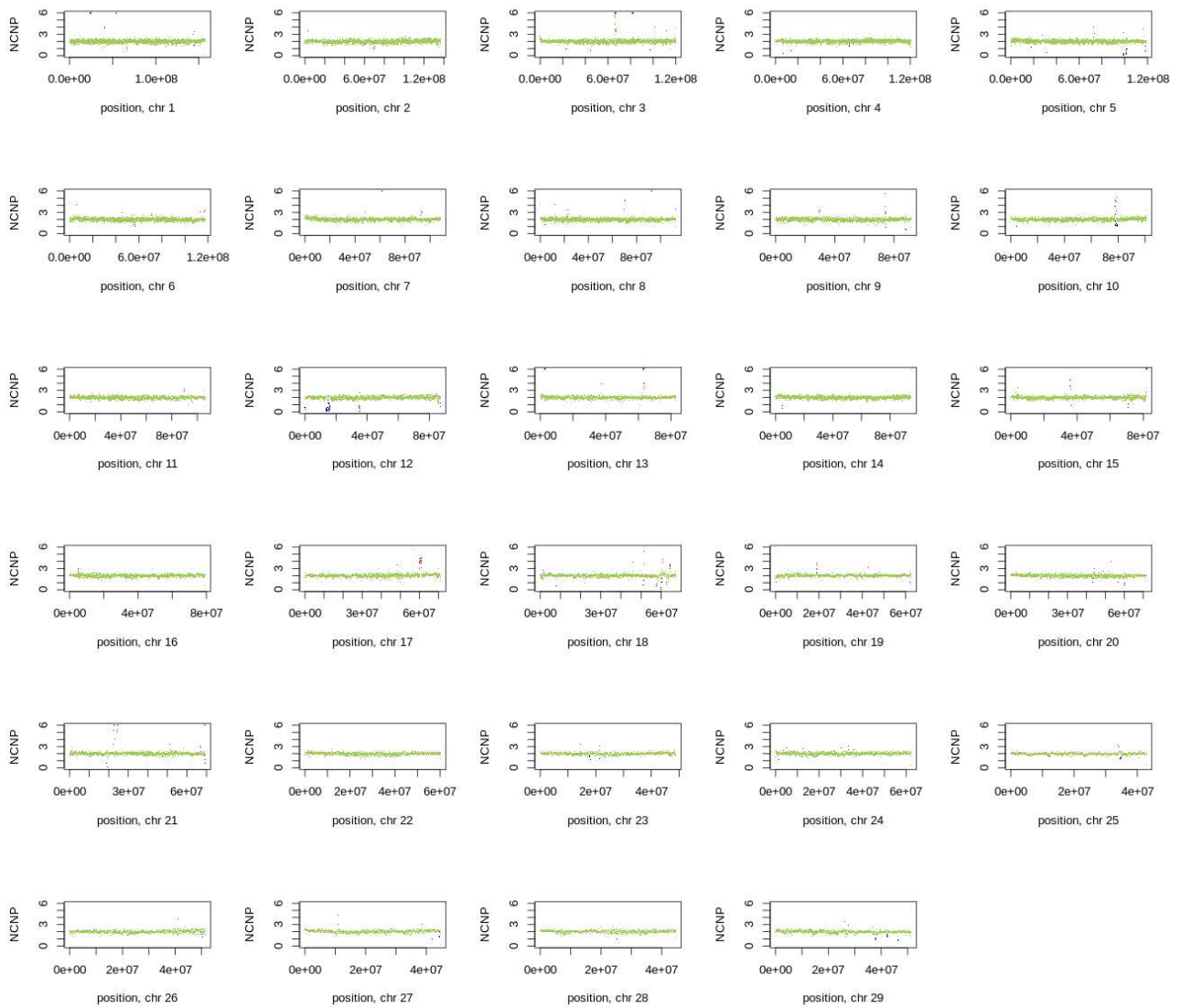


Figura 25: CNVs identificati nel genoma del campione FAZ_198

Si può notare come la distribuzione delle CNVs presenti all'interno di ogni singolo cromosoma sia simile nei due campioni. A titolo di esempio sotto in Figura 26 viene riportato il profilo delle CNVs per il cromosoma 10 caprino dei due campioni.

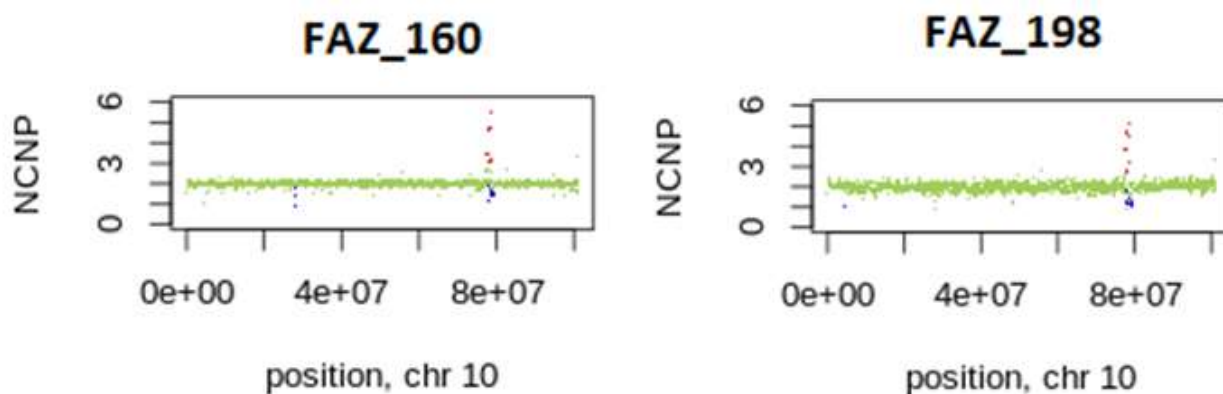


Figura 26: CNVs per il cromosoma 10 dei due campioni sequenziati.

Una volta individuate le CNVs attraverso i dati di NGS, queste sono state utilizzate per validare le CNVs individuate con l'analisi eseguite dai *genotyping arrays*.

Un totale di 92.475 CNVs sono state utilizzate per fare un *cross-match* con le 15 CNVRs precedentemente individuate dalla *consensus* analisi. Delle 15 CNVRs *consensus* sono state identificate e confermate con successo 5 di esse (Tabella 8). La lunghezza media totale delle CNVRs validate era di 647.527 bp con un *range* di lunghezza da 75.025 bp a 1.253.262 bp. Inoltre, la maggior parte delle CNVRs, ovvero l'80% di esse erano CNVRs *Gain*. Le 5 CNVRs validate erano presenti sui cromosomi 10, 15, 17, 18 e 24 (in rosso e grassetto all'interno della Tabella 8).

Cromosoma	Pos. iniziale	Pos. finale	Lunghezza	Status
7	88.779.618	88.952.154	172.537	LOSS
8	15.508.262	15.587.169	78.908	LOSS
8	32.560.844	32.672.600	111.757	LOSS
10	77.397.589	78.650.850	1.253.262	GAIN
13	3.796	252.595	248.800	GAIN
13	42.987.310	43.075.835	88.526	LOSS
14	26.164.070	26.259.929	95.860	LOSS
15	3.836.810	3.911.834	75.025	GAIN
15	22.129.446	22.216.521	87.076	GAIN
15	30.481.062	30.600.838	119.777	LOSS
16	18.319.045	18.428.267	109.223	LOSS
16	57.484.593	57.584.206	99.614	LOSS
17	60.084.864	61.156.582	1.071.719	GAIN
18	62.766.793	62.916.056	149.264	COMPLEX
24	59.774.288	60.462.652	688.365	LOSS

Tabella 8: Lista delle CNVRs *consensus* identificate e relativo status nei due software. In rosso e grassetto sono indicate le 5 CNVRs validate con NGS

Infine, l'aver confermato e validato 5 CNVRs delle 15 CNVRs *consensus*, sta a significare che il 33,33% del totale è stato validato; questo risultato rappresenta un valore molto più alto rispetto alle comuni procedure di validazione che si basano sul testare un esiguo numero di CNVRs rispetto al totale di quelle

identificate. In questo modo, le CNVRs identificate con i dati dei BeadChip sono state confrontate con quelle identificate con dati NGS sfruttando quella che è la maggiore risoluzione propria di questa tecnologia.

3 Annotazione genica

L'annotazione genica delle CNVRs *consensus* validate con il sequenziamento (Tabella 8), è stata effettuata con la sequenza di riferimento del genoma di capra GCF_001704415.1, anche nota come ARS1, attraverso l'utilizzo del browser Genome Data Viewer del National Center for Biotechnology Information Database (https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?context=genome&acc=GCF_001704415.1).

Sono stati trovati 84 geni all'interno delle 5 CNVRs validate. La maggior parte di essi erano presenti all'interno della CNVR posizionata sul cromosoma 10 (78,57%) nella regione 77.397.589 bp – 78.650.850 bp. Tra i diversi geni, 66 (78,57%) erano quelli responsabili della risposta immunitaria, 3 (3,57%) riguardanti i recettori olfattivi, 7 (8,33%) geni riguardavano le proteine di trasporto, 5 (5,95%) geni responsabili dei trasportatori degli amminoacidi cationici, uno (1,20%) riguardante la proteina ribonucleare e due (2,38%) riguardanti le caderine.

All'interno della CNVRs posizionata sul cromosoma 10 vi erano numerosi geni responsabili della risposta immunitaria: *T-cell receptor* o recettore dei linfociti T (o TCR). Questo è un recettore transmembrana che si trova sulla superficie dei linfociti T, responsabile del riconoscimento degli antigeni presentati dal complesso maggiore di istocompatibilità (MHC). Inoltre, tra i diversi geni erano presenti quelli delle immunoglobuline K (IgK). Entrambi sono coinvolti nello sviluppo o nel funzionamento del sistema immunitario, cioè un sistema di protezione degli organismi messo in atto per rispondere a potenziali minacce interne o invasive. Come dimostrato da Sasaki e coll. (2016) (Sasaki S. 2016) e da de Lemos e coll. (2018) (de Lemos M.V.A. 2018), attraverso la comparazione di geni presenti in altri mammiferi in particolare nel genoma di topo, si è evidenziato come questi geni responsabili della risposta immunitaria erano contenuti nelle CNVRs dei bovini ciò si evidenzia anche in questo lavoro in cui all'interno della CNVR presente sul cromosoma 10 e lunga 1.253.262 bp erano presenti questi geni. Alcuni autori hanno riportato che vi è la presenza di geni sovrarappresentati relativi alla risposta del sistema immunitario e sensoriale nei bovini e nell'uomo (Nguyen D.Q. 2006), così come è accaduto anche in questo lavoro. Tutto ciò suggerisce che il beneficio per la sopravvivenza può essere raggiunto con l'aumento del dosaggio di geni relativi al sistema immunitario e ambientale (Stothard P. 2011).

Nella CNVR presente sul cromosoma 15, sono stati trovati i geni LOC102184269, LOC102183988, LOC102183711, che sono responsabili della produzione del recettore olfattivo 1052-like (OR1052). Negli animali, i recettori chemosensoriali sono usati per trovare cibo, rilevare compagni e prole, riconoscere i territori ed evitare il pericolo come dimostrato nei bovini (Di Gerlando R. 2018), negli ovini (Liu J. 2013) e in altri mammiferi (Hughes G.M. 2018). Alcune variazioni nel numero di copie possono spiegare

l'adattamento degli organismi a diversi ambienti infatti le CNVs riguardanti i recettori olfattivi possono essere utilizzate per spiegare quello che può essere l'adattamento di una popolazione quando si sposta da un posto ad un altro oppure per differenziare quello che può essere l'adattamento di un animale in una nicchia sociale o quando conduce uno stile di vita solitario (Hughes G.M. 2018).

Sul cromosoma 17, in una CNVR lunga 1.071.719 bp sono stati ritrovati i geni EDNRA, TMEM184C, PRMT9, ARHGAP10, NR3C2. Questi stessi geni sono stati riscontrati anche nel lavoro di Menzi e coll. 2016 (Menzi F. 2016). Questi geni si trovano sul cromosoma 17, tra la posizione 60.084.864 e 61.156.582, in un punto diverso rispetto a quanto dichiarato da Menzi e coll. dovuto alla versione del genoma di riferimento usato. Questi geni svolgono un ruolo nella pigmentazione del colore del mantello, tra questi il gene NR3C2 codifica per il recettore dei mineralcorticoidi coinvolto nella regolazione della pressione arteriosa. Il gene ARHGAP10 codifica per la Rho GTPase activating protein 10, PRMT9 codifica per la proteina arginina metiltransferasi 9 e TMEM184C codifica per la proteina transmembrana 184C. EDNRA codifica per il recettore endotelina di tipo A coinvolto nello sviluppo dell'arco branchiale ma ancora non si sa se sia direttamente coinvolto nella pigmentazione. Inoltre, è stato dimostrato che l'aumento del numero di copie a questo CNV può portare a una riduzione della pigmentazione del mantello in modo dose-dipendente.

Altro gene trovato lungo questo cromosoma è il TRNAC-ACA il quale codifica per il tRNA-cisteina anticodone ACA il quale come dimostrato da Cecchi e coll. codifica per un tRNA-cisteina la cui funzione è quella di tradurre i codoni UGU e UGC negli mRNA ai residui di cisteina durante la sintesi proteica citoplasmatica (Cecchi F. 2017).

Sul cromosoma 18 sono stati trovati i geni LOC102191827, LOC102169326, LOC108633258, LOC102191076, LOC102183956, i quali codificano per i trasportatori 3 degli amminoacidi cationici (CAT3). I membri della famiglia CAT trasportano essenzialmente amminoacidi cationici mediante diffusione facilitata con trans-stimolazione differenziale da substrati intracellulari. In alcune cellule, possono regolare il tasso di sintesi di ossido nitrico (NO) controllando l'assorbimento di l-arginina come substrato per l'ossido nitrico sintasi (NOS). È stato dimostrato in altri mammiferi, in particolare uomo e ratto, che questo recettore è localizzato esclusivamente nei neuroni centrali. In umano, l'espressione di CAT-3 è presente nel cervello anche se l'espressione più elevata è stata trovata nel timo (Vékony N. 2001). Nell'ultima CNVR validata, posizionata sul cromosoma 24, sono stati trovati i geni LOC102176484, CDH20, LOC108633808, presenti tra la posizione 59.774.288 e la posizione 60.462.652. I prodotti che vengono codificati da questi geni sono *heterogeneous nuclear ribonucleoprotein H-like pseudogene*, e *cadherin 20*, rispettivamente per LOC102176484, CDH20, LOC108633808.

Le ribonucleoproteine nucleari eterogenee (hnRNP) comprendono una famiglia di proteine leganti l'RNA. La complessità e la diversità associate agli hnRNP li rendono multifunzionali, e queste sono coinvolte non solo nell'elaborazione di RNA nucleari eterogenei (hnRNA) in maturi mRNA, ma anche come trans-fattori nella regolazione dell'espressione genica. Le proteine hnRNP influenzano la struttura degli hnRNA e facilitano o ostacolano l'interazione delle sequenze degli hnRNA con altri componenti necessari per

l'elaborazione dei pre-mRNA, influenzando così il destino degli hnRNA. Le hnRNP possono svolgere anche ruoli importanti nell'interazione degli hnRNA con altre strutture nucleari, nel trasporto nucleocitoplasmatico degli mRNA e in altri processi cellulari. Le proteine hnRNP sono abbondanti nella crescita delle cellule vertebrate in particolare a livello degli istoni e dei complessi di proteine hnRNA-hnRNP (hnRNP complex) (Dreyfuss G. 1993), tra queste proteine hnRNP l'isoforma H sembra funzionare come fattore di cleavage/poliadenilazione intervenendo quindi in modo significativo sia questa isoforma che le altre sull'espressione genica.

Tra gli altri geni, all'interno della stessa regione, sono state trovate le caderine 20. Le caderine sono proteine di adesione cellulare calcio-dipendenti che interagiscono preferenzialmente con loro stesse nel connettere le cellule. È stato riscontrato da Wiech e coll. come una CNV loss nel gene CDH20 possa essere un possibile punto di partenza del coinvolgimento delle molecole di adesione nella progressione dell'adenocarcinoma esofageo (Wiech T. 2009).

È stato visto che durante lo sviluppo del midollo spinale del pollo, la caderina 20 viene espressa dai motoneuroni della colonna motoria laterale e da altre cellule nelle placche basali e alari (Luo J. 2009). La tendenza delle caderine per accumulare CNVs, può riferirsi alla loro struttura altamente ripetitiva contenente caderina, laminina, ripetizioni A e G, EGF e mucina che possono risultare una fonte di instabilità genomica (Seroussi E. 2010). Attraverso studi di associazione *genome-wide* negli ultimi anni è stato visto come le caderine siano coinvolte e possano essere dei potenziali geni candidati per il contenuto di colesterolo nel latte bovino (Do D.N. 2018). Si è visto anche che vi può essere una probabile associazione di queste con quella che è la variazione delle caratteristiche delle fibre in particolare si è visto come possano essere coinvolte nella determinazione del fenotipo lanoso nelle fibre di capra (Li Y. 2018). Quindi come si evince molti di questi geni possono essere coinvolti in quelli che sono i caratteri di interesse economico, relativi alla produzione ma anche riguardanti le funzioni del sistema immunitario ed i meccanismi regolatori interni.

Capitolo IV: Conclusioni

Nonostante nel corso degli anni passati, sono stati fatti alcuni studi sulle CNVs nella specie caprina, il presente studio è uno dei pochi ad utilizzare un'analisi dei dati provenienti da SNP per rilevare CNVs in questa specie, in particolare nella razza caprina Girgentana, utilizzando due diversi algoritmi e validandoli attraverso dati provenienti da metodiche superiori come l'NGS. L'utilizzo del BeadChip Illumina GoatSNP50, a media densità, potrebbe portare nel complesso all'individuazione di falsi positivi, per cui l'utilizzo delle nuove tecnologie di sequenziamento per validare la chiamata delle CNVs può portare al superamento di questi limiti. Inoltre, l'utilizzo di due differenti software, basati su due differenti algoritmi, ha evidenziato la già nota variabilità dei risultati ottenuti ed ha reso necessaria una prima individuazione di CNVRs *consensus* comuni ad entrambi. Attraverso il sequenziamento NGS si sono riscontrati un numero di CNVs maggiore rispetto a quelli che si possono trovare con i BeadChip. Inoltre, la validazione dei risultati tramite il sequenziamento individuale, data la maggiore risoluzione propria di questa tecnologia, ha permesso di validare un numero di CNVRs altamente soddisfacente (33,33%).

Infine, molti geni rilevati all'interno delle CNVRs validate hanno ruoli importanti nell'adattamento o nella resistenza alle malattie, su alcuni caratteri produttivi e nei meccanismi regolatori interni.

Pertanto i nostri risultati hanno fornito delle informazioni importanti andando a validare un numero superiore di CNVs totali rispetto a quanto fatto in altri studi, il tutto per la costruzione di una mappa più completa delle CNVs del genoma caprino. Questo lavoro, inoltre offre anche una risorsa importante per gli studi futuri che verranno fatti su questa specie per arricchire ancora di più il genoma di questo animale poco studiato.

Bibliografia

- (MiPAAF), Ministero delle Politiche Agricole Alimentari e Forestali. «Linee guida per la conservazione e la caratterizzazione della biodiversità animale di interesse per l'agricoltura.» A cura di Baldassarre Portolano, Alessio Zanon Riccardo Fortina. 2012.
- B.J. Norris, V.A. Whan. «A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep.» *Genome Research* 18 (2008): 1282-1293.
- Bae J.S., Cheong H.S., Kim L.H., NamGung S., Park T.J., Chun J.Y., Kim J.Y., Pasaje C.F., Lee J.S., Shin H.D. «Identification of copy number variations and common deletion polymorphisms in cattle.» *BMC genomics* 11 (2010): 232.
- Boeva V., Popova T., Bleakley K., Chiche P., Cappo J., Schleiermacher G., Janoueix-Lerosey I., Delattre O., Barillot E. «Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.» *Bioinformatics* 28 (2011): 423-425.
- Buckland P.R.. «Polymorphically duplicated genes: their relevance to phenotypic variation in humans.» *Annals of medicine* 35 (2003): 308-315.
- Cañón J., García D., García-Atance M.A., Obexer-Ruff G., Lenstra J.A., Ajmone-Marsan P., Dunner S., The ECONOGENE Consortium. «Geographical partitioning of goat diversity in Europe and the Middle East.» *Animal Genetics* 37 (2006): 327-334.
- Capre.it. <https://www.capre.it/le-razze-in-italia/registro-anagrafico-caprino/12-girgentana.html>.
- Carreto L., Eiriz M.F., Gomes A.C., Pereira P.M., Schuller D., Santos M.A.S. «Comparative genomics of wild type yeast strains unveils important genome diversity.» *BMC Genomics* 9 (2008): 524.
- Cecchi F., Russo C., Iamartino D., Galiero A., Turchi B., Fratini F., Degl'Innocenti S., Mazza R., Biffani S., Preziuso G., Cantile C. «Identification of candidate genes for paratuberculosis resistance in the native Italian Garfagnina goat breed.» *SpringerLink* 49 (2017): 1135-1142.
- Chen W.K., Swartz J.D., Rush L.J., Alvarez C.E. «Mapping DNA structural variation in dogs.» *Genome research* 19 (2009): 500-509.
- Chen Y., Zhao L., Wang Y., Cao M., Gelowani V., Xu M., Agrawal S.A., Li Y., Daiger S.P., Gibbs R., Wang F., Chen R. «SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data.» *BMC Bioinformatics* 18 (2017): 147.
- Clop A., Vidal O., Amills M. «Copy number variation in the genomes of domestic animals.» *Animals Genetics* 43 (2012): 503-517.
- Conrad D.F., Andrews T.D., Carter N.P., Hurles M.E., Pritchard J.K. «A high-resolution survey of deletion polymorphism in the human genome.» *Nature Genetics* 38 (2006): 75-81.
- Cornale P., Renna M., Lussiana C., Bigi D., Chessa S., Mimosi A. «The Grey Goat of Lanzo Valleys (Fiurinà): breed characteristics, genetic diversity, and quantitative-qualitative milk traits.» *Small Ruminant Research* 116 (2014): 1-13.

- da Silva J.M., Giachetto P.F., da Silva L.O., Cintra L.C., Paiva S.R., Yamagishi M.E.B., Caetano A.R. «Genome-wide copy number variation (CNV) detection in Nelore cattle reveals highly frequent variants in genome regions harboring QTLs affecting production traits.» *BMC Genomics* 17 (2016).
- de Lemos M.V.A., Berton M.P., de Camargo G.M.F., Peripolli E., de Oliveira Silva R.M., Olivieri B.F., Cesar A.S.M., Pereira A.S.C., de Albuquerque L.G., de Oliveira H.N., Tonhati H., Baldi F. «Copy number variation regions in Nellore cattle: Evidences of environment adaptation.» *Livestock Science* 207 (2018): 51-58.
- DeRisi J., Penland L., Brown P.O., Bittner M.L., Meltzer P.S., Ray M., Chen Y., Su Y.A., Trent J.M.. «Use of a cDNA microarray to analyse gene expression patterns in human cancer.» *Nature genetics* 14 (1996): 457-460.
- Di Gerlando R., Sardina M.T., Tolone M., Sutera A.M., Mastrangelo S., Portolano B. «Genome-wide detection of copy-number variations in local cattle breeds.» *Animal Production Science*, 2018.
- Di Gerlando R., Tortorici L., Sardina M.T., Monteleone G., Mastrangelo S., Portolano B. «Molecular characterisation o k-casein gene in Girgentana dairy goat breed and identification of two new alleles.» *Italian Journal of Animal Science* 14 (2015): 90-93.
- Do D.N., Schenkel F.S., Miglior F., Zhao X., Ibeagha-Awemu E.M. «Genome wide association study identifies novel potential candidate genes for bovine milk cholesterol content.» *Scientific reports* 8 (2018).
- Dong Y., Zhang X., Xie M., Arefnezhad B., Wang Z., Wang W., Feng S., Huang G., Guan R., Shen W., Bunch R., McCulloch R., Li Q., Li B., Zhang G., Xu X., Kijas J.W., Salekdeh G.H., Wang W., Jiang Y. «Reference genome of wild goat (*capra aegagrus*) and sequencing of goat breeds provide insight into genic basis of goat domestication.» *BMC genomics* 16 (2015).
- Dreyfuss G., Matunis M.J., Piiiol-Roma S., Burd C.G. «hnRNP Proteins and the biogenesis of mRNA.» *Annual Reviewes Biochemistry* 62 (1993): 289-321.
- Emerson J.J., Cardoso-Moreira M., Borevitz J.O., Long M. «Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*.» *Science* 320 (2008): 1629-1631.
- Fadista J., Thomsen B., Holm L.E., Bendixen C. «Copy number variation in the bovine genome.» *BMC Genomics* 11, n. 284 (2010).
- Fanciulli M., Norsworthy P.J., Petretto E., Dong R., Harper L., Kamesh L., Heward J.M., Gough S.C., de Smith A., Blakemore A.I., Froguel P., Owen C.J., Pearce S.H., Teixeira L., Guillevin L., Graham D.S., Pusey C.D., Cook H.T., Vyse T.J., Aitman T.J. «FCGR3B copy number variation is associated with susceptibility to systemic, but not organspecific autoimmunity.» *Nature genetics* 39 (2007): 721-723.
- FAO. «The state of the world's animal genetic resources for food and agriculture. Commission on genetic resources for food and agriculture.» Rome: FAO, 2007.
- Fatet A., Pellicer-Rubio M.T., Leboeuf B. «Reproductive cycle of goats.» *Animal Reproduction Science* 124 (2011): 211-219.

- Fontanesi L., Beretti F., Martelli P.L., Colombo M., Dall'Olio S., Occidente M., Portolano B., Casadio R., Matassino D., Russo V. «A first comparative map of copy number variations in the sheep genome.» *Genomics*, 2011.
- Fontanesi L., Beretti F., Riggio V., Gómez González E., Dall'Olio S., Davoli R., Russo V., Portolano B. «Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors.» *Cytogenetic and genome research* 126 (2009): 333-347.
- Fontanesi L., Martelli P.L., Beretti F., Riggio V., Dall'Olio S., Colombo M., Casadio R., Russo V., Portolano B. «An initial comparative map of copy number variations in the goat (*Capra hircus*) genome.» *BMC Genomics* 11 (2010).
- Graubert T.A., Cahan P., Edwin D., Selzer R.R., Richmond T.A., Eis P.S., Shannon W.D., Li X., McLeod H.L., Cheverud J.M., Ley T.J. «A high-resolution map of segmental DNA copy number variation in the mouse genome.» *Plos Genetics* 3 (2007).
- Griffin D.K., Robertson L.B., Tempest H.G., Vignal A., Fillon V., Crooijmans R.P., Groenen M.A., Deryusheva S., Gaginskaya E., Carré W., Waddington D., Talbot R., Völker M., Masabanda J.S., Burt D.W. «Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution.» *BMC Genomics* 9 (2008): 168.
- Groeneveld L.F., Lenstra J.A., Eding H., Toro M.A., Scherf B., Pilling D., Negrini R., Finlay E.K., Jianlin H., Groeneveld E., Weigend S., GLOBALDIV Consortium. «Genetic diversity in farm animals--a review.» *Animal Genetics* 41 (2010): 6-31.
- Gurgul A., Semik E., Pawlina k., Szmatoła T., Jasielczuk I., Bugno-Poniewierska M. «The application of genome-wide SNP genotyping methods in studies on livestock genomes.» *journal of applied genetics* 55 (2014): 197-208.
- Heaton M.P., Keen J.E., Clawson M.L., Harhay G.P., Bauer N., Shultz C., Green B.T., Durso L., Chitko-McKown C.G., Laegreid W.W. «Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing.» *Journal of the American Veterinary Medical Association* 226 (2005).
- Henrichsen C.N., Chaignat E., Reymond A. «Copy number variants, diseases and gene expression.» *Human molecular genetics* 18 (2009): R1-R8.
- Hou Y., Liu G.E., Bickhart D.M., Cardone M.F., Wang K., Kim E.S., Matukumalli L.K., Ventura M., Song J., VanRaden P.M., Sonstegard T.S., Van Tassell C.P. «Genomic characteristics of cattle copy number variations.» *BMC genomics* 12 (2011): 127.
- Huang J., Li R., Zhang X., Huang y., Dang R., Lan X., Chen H., Lei C. «Copy number variation regions detection in Qinchuan cattle.» *Livestock Science (Elsevier)* 204 (2017): 88-91.
- Hughes G.M., Boston E.S.M., Finarelli J.A., Murphy W.J., Higgins D.G., Teeling E.C. «The Birth and Death of Olfactory Receptor Gene Families in Mammalian Niche Adaptation.» *Molecular Biology and Evolution* 35 (2018): 1390-1406.
- Illumina. s.d. <https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/nextera-xt-troubleshooting-technical-note.pdf>.

- Jenkins G.M., Goddard M.E., Black M.A., Brauning R., Auvray B., Dodds K.G., Kijas J.W, Cockett N., McEwan J.C. «Copy number variants in the sheep genome detected using multiple approaches.» *BMC Genomics* 17 (2016).
- Jiang L., Jiang J., Yang J., Liu X., Wang J., Wang H., Ding X., Liu J. «Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins.» *BMC Genomics* 14 (2013).
- Jordan S. A., Humpries P. «Single nucleotide polymorphism in exon 2 of the BCP gene on 7q31-q35.» *Human Molecular Genetics* 3 (1994).
- Lee A.S., Gutiérrez-Arcelus M., Perry G.H., Vallender E.J., Johnson W.E., Miller G.M., Korbel J.O., Lee C. «Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies.» *Human Molecular Genetics* 17 (2008): 1127-1136.
- Li Y., Zhou G., Zhang R., Guo J., Li C., Martin G., Chen Y., Wang X. «Comparative proteomic analyses using iTRAQ-labeling provides insights into fiber diversity in sheep and goats.» 172 (2018): 82-88.
- Linton Y., Harbach R.E., Seng C.M., Anthony T.G., Matusop A. «Morphological and molecular identity of Anopheles (Cellia) sundaicus (Diptera: Culicidae), the nominotypical member of malaria vector species complex in Souteast Asia.» *Systematic Entomology* 26 (2001): 357-366.
- Liu G.E., Hou Y., Cardone M.F., Jiang L., Cellamare A., Mitra A., Alexander L.J., Coutinho L.L., Dell'Aquila M.E., Gasbarre L.C., Lacalandra G., Li R.W., Matukumalli L.K., Nonneman D., Regitano L.C., Smith T.P., Song J., Sonstegard T.S., Van Tassel C.P., Ventura M., Eichler E.E., McDanel T.G., Keele J.W. «Analysis of copy number variations among diverse cattle breeds.» *Genome research* 20 (2010): 693-703.
- Liu G.E., Van Tassel C.P., Sonstegard T.S., Li R.W., Alexander L.J., Keele J.W., Matukumalli L.K., Smith T.P., Gasbarre L.C. «Detection of germline and somatic copy number variations in cattle.» *Developments in biologicals* 132 (2008): 231-237.
- Liu J., Zhang L., Xu L., Ren H., Lu J., Zhang X., Zhang S., Zhou X., Wei C., Zhao F., Du L. «Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array.» *BMC Genomics* 14 (2013).
- Lucito R., Healy J., Alexander J., Reiner A., Esposito D., Chi M., Rodgers L., Brady A., Sebat J., Troge J., West J.A., Rostan S., Nguyen K.C., Powers S., Ye K.Q., Olshen A., Venkatraman E., Norton L., Wigler M. «Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation.» *Genome research* 13 (2003): 2291-2305.
- Luo J., Ju Min J., Lin J., Yan X., Markus A., Mix E., Rolfs A., Redies C. «Cadherin-20 expression by motor neurons is regulated by Sonic hedgehog during spinal cord development.» *NeuroReport* 20 (2009): 365-370.
- Ma Y., Zhang Q., Lu Z., Zhao X., Zhang Y. «Analysis of copy number variations by SNP50 BeadChip array in Chinese sheep.» *Genomics* 106 (2015): 295-300.

- Mannen H., Nagata Y., Tsuji S. «Mitochondrial DNA Reveal that Domestic Goat (*Capra hircus*) are Genetically Affected by Two Subspecies of Bezoar (*Capra aegagurus*).» *Biochemical Genetics* 39 (Giugno 2001): 145-154.
- Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L., Jarvie T.P., Jirage K.B., Kim J.B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F., Rothberg J.M. «Genome sequencing in microfabricated high-density picolitre reactors.» *Nature* 437 (2005): 376-380.
- Mastrangelo S., Tolone M., Montalbano M., Tortorici L., Di Gerlando R., Sardina M.T., Portolano B. «Population genetic structure and milk production traits in Girgentana goat breed.» *Animal Production Science*, Marzo 2016.
- Maydan J.S., Lorch A., Edgley M.L., Flibotte S., Moerman D.G. «Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*.» *BMC Genomics* 11 (2010): 62.
- McCarroll S.A., Hadnott T.N., Perry G.H., Sabeti P.C., Zody M.C., Barrett J.C., Dallaire S., Gabriel S.B., Lee C., Daly M.J., Altshuler D.M., Consortium International HapMap. «Common deletion polymorphisms in the human genome.» *Nature Genetics* 38 (2006): 86-92.
- Meadow R.H.. *Animal domestication in the Middle East: a revised view from the eastern margin. In: Harappan Civilization*. A cura di G.I. Possehl. New Delhi: Oxford & IBH, 1993.
- Menzi F., Keller I., Reber I., Beck J., Brenig B., Schütz E., Leeb T., Drögemüller C. «Genomic amplification of the caprine EDNRA locus might lead to a dose dependent loss of pigmentation.» *Scientific Reports* 6, n. 28438 (2016).
- Miller N.J., Guillemaud T., Giordano R., Siegfried B.D., Gray M.E., Meinke L.J., Sappington T.W. «Genes, gene flow and adaption of *Diabrotica virgifera virgifera*.» *Agricultural and Forest Entomology* 11 (2009): 47-60.
- Negrini R., Nicoloso L., Crepaldi P., Milanesi E., Marino R., Pariset L., Leveziel H., Williams J.L., Marsan P.A. «Traceability of four European Protected Geographic Indication (PGI) beef products using Single Nucleotide Polymorphisms(SNP) and Bayesian statistics.» *Meat Science*, 2008.
- Nguyen D.Q., Webber C., Ponting C.P. «Bias of Selection on Human Copy-Number Variants.» *Plos Genetics* 2 (2006).
- Nicoloso L., Bomba L., Colli L., Negrini R., Milanesi M., Mazza R., Sechi T., Frattini S., Talenti A., Coizet B., Chessa S., Marletta D., D'Andrea M., Bordonaro S., Ptak G., Carta A., Pagnacco G., Valentini A., Pilla F., Ajmone-Marsan P., Crepaldi P., Italian Goat Consortium. «Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip.» *Genetics, selection, evolution* 47 (2015).
- Norris B.J., Whan V.A. «A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep.» *Genome Research* 18 (2008): 1282-1293.

- Ossowski S., Schneeberger K., Clark R.M., Lanz C., Warthmann N., Weigel D. «Sequencing of natural strains of *Arabidopsis thaliana* with short reads.» *Genome Research* 18 (2008): 2024-2033.
- Parejo J.C., Padilla J.A., Calero R., Martínez-Trancón M., Sansinforiano E., Rabasco A., Salazar J., Bravo J.A. «Population genetic structure and conservation management of Retinta Extremeña goats.» *Small Ruminant Research* 124 (2015): 9-16.
- Pelotti S., Ceccardi S., Alù M., Lugaresi F., Trane R., Falconi M., Bini C., Ciccogni A. «Cancerous tissues in forensic genetic analysis.» *Genet Test* 11 (2007): 397-400.
- PennCNV*. <http://penncnv.openbioinformatics.org/en/latest/>.
- Perry G.H., Yang F., Marques-Bonet T., Murphy C., Fitzgerald T., Lee A.S., Hyland C., Stone A.C., Hurles M.E., Tyler-Smith C., Eichler E.E., Carter N.P., Lee C., Redon R. «Copy number variation and evolution in humans and chimpanzees.» *Genome Research* 18 (2008): 1698-1710.
- Pierce M.D., Dzama k., Muchadeyi F.C. «Genetic Diversity of Seven Cattle Breeds Inferred Using Copy Number Variations.» *Frontier in Genetics* 9 (2018).
- Poetsch M., Petersmann A., Woenckhaus C., Protzel C., Dittberner T., Lignitz E., Kleist B. «Evaluation of allelic alterations in short tandem repeats in different kinds of solid tumors-possible pitfalls in forensic casework.» *Forensic Science International* 145 (2004): 1-6.
- Popova T., Manié E., Stoppa-Lyonnet D., Rigail G., Barillot E., Stern M.H. «Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays.» *Genome biology* 10 (2009): R128.
- Prinsen R.T.M.M., Rossoni A., Gredler B., Bieber A., Bagnato A., Strillacci M.G. «A genome wide association study between CNVs and quantitative traits in Brown Swiss cattle.» *Livestock Science* 202 (2017): 7-12.
- Prinsen R.T.M.M., Strillacci M.G., Schiavini F., Santus E., Rossoni A., Maurer V., Bieber A., Gredler B., Dolezal M., Bagnato A. «A genome-wide scan of copy number variants using high-density SNPs in Brown Swiss dairy cattle.» *Livestock Science* 191 (2016).
- Purfield D.C., Berry D.P., McParland S., Bradley D.G. «Runs of homozygosity and population history in cattle.» *BMC Genetics* 13 (2012): 70.
- Quinlan A.R., Hall I.M. «BEDTools: a flexible suite of utilities for comparing genomic features.» *Bioinformatics* 26 (2010): 841-842.
- Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R., Chen W., Cho E.K., Dallaire S., Freeman J.L., González J.R., Gratacòs M., Huang J., Kalaitzopoulos D., Komura D., MacDonald J.R., Marshall C.R., Mei R., Montgomery L., Nishimura K., Okamura K., Shen F., Somerville M.J., Tchinda J., Valsesia A., Woodwark C., Yang F., Zhang J., Zerjal T., Zhang J., Armengol L., Conrad D.F., Estivill X., Tyler-Smith C., Carter N.P., Aburatani H., Lee C., Jones K.W., Scherer S.W., Hurles M.E. «Global variation in copy number in the human genome.» *Nature* 444 (2006): 444-454.
- Sardina M.T., Ballester M., Marmi J., Finocchiaro R., van Kaam J.B.C.H.M., Portolano B. «Phylogenetic analysis of Sicilian goats reveals a new mtDNA lineage.» *Journal compilation*, 2006: 376-378.

- Sasaki S., Watanabe T., Nishimura S., Sugimoto Y. «Genome-wide identification of copy number variation using high-density singlenucleotide polymorphism array in Japanese Black cattle.» *BMC Genetics* 17 (2016).
- Schaschl H., Aitman T.J., Vyse T.J. «Copy number variation in the human genome and its implication in autoimmunity.» *Clinical and experimental immunology* 156 (2009): 12-16.
- Schena M., Shalon D., Davis R.W., Brown P.O. «Quantitative monitoring of gene expression patterns with a complementary DNA microarray.» *Science* 270 (1995): 467-470.
- Schrider D.R., Hahn M.W. «Gene copy-number polymorphism in nature.» *Biological sciences* 277 (2010): 3213-3221.
- Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Månér S., Massa H., Walker M., Chi M., Navin N., Lucito R., Healy J., Hicks J., Ye K., Reiner A., Gilliam T.C., Trask B., Patterson N., Zetterberg A., Wigler M. «Large-scale copy number polymorphism in the human genome.» *Science* 305 (2004): 525-528.
- Seroussi E., Glick G., Shirak A., Yakobson E., Weller J.L., Ezra E., Zeron Y. «Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs.» *BMC Genomics* 11, n. 673 (2010).
- She X., Cheng Z., Zöllner S., Church D.M., Eichler E.E. «Mouse segmental duplication and copy number variation.» *Nature Genetics* 40 (2008): 909-914.
- SNP & VARIATION SUITE. http://goldenhelix.com/products/SNP_Variation/index.html.
- Springer N.M., Ying K., Fu Y., Ji T., Yeh C.T., Jia Y., Wu W., Richmond T., Kitzman J., Rosenbaum H., Iniguez A.L., Barbazuk W.B., Jeddeloh J.A., Nettleton D., Schnable P.S. «Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content.» *PLoS Genetics* 5 (2009).
- Stothard P., Choi J.L., Basu U., Sumner-Thomson J.M., Meng Y., Liao X., Moore S.S. «Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery.» *BMC Genomics* 12 (2011): 559.
- Vahidi S.M.F., Tarang A.R., Naqvi A.U.N., Falahati A.M., Boettcher P., Joost S., Colli L., Garcia J.F., Ajmone-Marsan P. «Investigation of the genetic diversity of domestic *Capra hircus* breeds reared within an early goat domestication area in Iran.» *Genetics, Selection, Evolution* 46 (2014): 27.
- Vékony N., Wolf S., Boissel J.P., Gnauert K., Closs E.I. «Human Cationic Amino Acid Transporter hCAT-3 Is Preferentially Expressed in Peripheral Tissues.» *Biochemistry* 40 (2001): 12387-12394.
- Wain L.V., Armour J.A., Tobin M.D. «Genomic copy number variation, human health, and disease.» *Lancet* 374 (2009): 340-350.
- Wallace L.M., Garwick S.E., Mei W., Belayew A., Coppee F., Ladner K.J., Guttridge D., Yang J., Harper S.Q. «DUX4, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo.» *Annals of neurology* 69 (2010).

- Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F., Hakonarson H., Bucan M. «PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.» *Genome Research* 17 (2007): 1665-1674.
- Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F.A., Hakonarson H., Bucan M. «PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.» *Genome Research* 17 (2007): 1665-1674.
- Wang M.D., Dzama K., Hefer C.A., Muchadeyi F.C. «Genomic population structure and prevalence of copy number variations in South African Nguni cattle.» *BMC Genomics* 16 (2015).
- Wiech T., Nikolopoulos E., Weis R., Langer R., Bartholomé K., Timmer J., Walch A.K., Höfler H., Werner. «Genome-wide analysis of genetic alterations in Barrett's adenocarcinoma using single nucleotide polymorphism arrays.» *Laboratory Investigation* 89 (2009): 385-397.
- Yan J., Blair H.T., Liub M., Li W., He S., Lei Chen, Dittmer K.E., Garrick D.J., Biggs P.J., Dukkupati V.S.R. «Genome-wide detection of autosomal copy number variants in several sheep breeds using Illumina OvineSNP50 BeadChips.» *Small Ruminant Research* 55 (2017): 24-32.
- Yang M., Lv J., Zhang L., Li M., Zhou Y., Lan X., Lei C., Chen H. «Association study and expression analysis of CYP4A11 gene copy number variation in Chinese cattle.» *Scientific Reports* 7 (2017).
- Zeder M.A., Hesse B. «The Initial Domestication of Goats (*Capra hircus*) in the Zagros Mountains 10,000 Years Ago.» *Science* 287 (Marzo 2000): 2254-2257.
- Zhang B., Chang L., Lan Y.X., Asif N., Guan F.L., Fu K.D., Li B., Yan X.C., Zhang B.H., Zhang Y.X., Huang Z.Y., Chen H., Yu J., Li B.S. «Genome-wide definition of selective sweeps reveals molecular evidence of trait-driven domestication among elite goat (*Capra species*) breeds for the production of dairy, cashmere, and meat.» *Gigascience*, 2018.
- Zhu C., Fan H., Zehu Yuan, Hu S., Ma X., Xuan J., Wang H., Zhang L., Wei C., Zhang Q., Zhao F., Du L. «Genome-wide detection of CNVs in Chinese indigenous sheep with different types of tails using ovine high-density 600K SNP arrays.» *Scientific reports* 6 (2016): 1-9.