Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN

BIOLOGIA EVOLUZIONISTICA E AMBIENTALE

CICLO XXVIII

COORDINATORE Prof. Barbujani Guido

Nuove metodologie in genomica di popolazioni, e applicazioni a dati reali

Settore Scientifico Disciplinare BIO/18

Dottorando

Dott. Panziera Alex

Tutore

Prof. Bertorelle Giorgio

Anni 2013/2015

Nuove metodologie in genomica di popolazioni, e applicazioni a dati reali

Indice

Inquadramento della tesi Pag.1

Introduzione Pag.3

Le tecnologie di sequenziamento di nuova generazione Pag.3

Le tecnologie di sequenziamento di terza generazione Pag.5

Le sotto-rappresentazioni del genoma Pag.7

Inferenze evolutive su dati genomici Pag.8

La ricerca di loci soggetti al processo di selezione Pag.9

4P: Parallel Processing of Polymorphism Panels Pag.13

Introduzione Pag.13

Caratteristiche del software Pag.15

Performance e confronto con altre applicazioni Pag.17

Conclusione Pag.22

Glossario Pag.22

Determinanti genetiche e domesticazione: il caso del fagiolo comune (*Phaseolus vulgaris*). Pag.24

L'identificazione di geni sotto selezione nel fagiolo comune mesoamericano con dati RNA-seq Pag.25

Introduzione Pag.25

Materiali e metodi Pag.28

Risultati Pag.35

Discussione e conclusioni Pag.40

Materiali supplementari Pag.44

Analisi di selezione in un dataset ristretto di geni a nota omologia Pag.46

Materiali e metodi Pag.46

Risultati Pag.50

Discussione Pag.55

Conclusione Pag.57

Materiali supplementari Pag.59

Il cambiamento della modalità riproduttiva in Zootoca vivipara Pag.61

Introduzione Pag.62

Materiali e metodi Pag.66

Risultati Pag.70

Discussione Pag.73

Conclusioni Pag.76

Materiali supplementari Pag.78

Il sequenziatore portatile di terza generazione MinION basato su nanopori: test per lo studio di loci MHC nel camoscio alpino Pag.87

Materiali e metodi Pag.88

Risultati Pag.93

Discussione Pag.97

Conclusioni Pag.99

Considerazioni finali Pag.100

Bibliografia Pag.102

Articoli pubblicati, sottomessi o in preparazione Pag.122

Finanziamento del progetto di dottorato Pag.124

Inquadramento della tesi

Lo sviluppo delle tecnologie di sequenziamento di nuova generazione negli ultimi dieci anni ha profondamente rivoluzionato tutti i rami della genetica, inclusa la genetica di popolazioni. La possibilità di produrre una grande quantità di dati a prezzi ragionevoli ha cambiato l'oggetto degli studi genetici, spostatosi da pochi marcatori a interi genomi (es. (Huang et al. 2012; S. Liu et al. 2014)) o sotto-rappresentazioni del genoma (es. (Hohenlohe et al. 2010b)). Lo sviluppo di nuove tecnologie di sequenziamento è un processo ancora in corso, e le tecnologie di sequenziamento di seconda generazione, il motore di questa decade di forte sviluppo (Van Dijk et al. 2014), sono oggi incalzate da una nuova generazione di tecnologie, che promettono una lunghezza maggiore delle letture, una minore quantità di materiale di partenza (approccio a singola molecola) e costi inferiori (Schadt, Turner, and Kasarskis 2010). Esempi importanti di queste nuove tecnologie, che fanno parte cosiddetta terza generazione, sono le tecnologie Pacific Biosciences SMRT e Oxford Nanopore.

Come tutte le innovazioni tecnologiche, i dati genomici prodotti da tecnologie di nuova generazione richiedono lo sviluppo di nuove metodologie per la loro analisi. Ad esempio, quando i genomi sono sequenziati in organismi non modello (dove mancano informazioni sui polimorfismi nella specie) o con un basso livello di copertura, vi è una forte incertezza legata alla chiamata dei genotipi. Questo problema ha portato per esempio allo sviluppo di ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014), una suite di software che permette di tenere in considerazione questa incertezza stimando le verosimiglianze dei genotipi dalle letture dirette del genoma. L'enorme quantità di informazione genomica richiede anche lo sviluppo di metodi specifici per l'inferenza della storia demografica (Schraiber and Akey 2015) o per la ricerca di geni o regioni del genoma implicati in processi adattativi (Oleksyk, Smith, and O'Brien 2010),(Vitti, Grossman, and Sabeti 2013). Questi nuovi metodi prendono in considerazione caratteristiche dell'intero genoma, come lo spettro delle frequenze alleliche (Gutenkunst et al. 2009; Laurent Excoffier et al. 2013), o la densità locale dei siti eterozigoti (H. Li and Durbin 2011; Schiffels and Durbin 2014), o sono in grado di individuare alleli che hanno aumentato la loro frequenza così velocemente che l'associazione a lungo raggio con i polimorfismi vicini non è stata ancora erosa dalla

ricombinazione (Pardis C. Sabeti et al. 2007). La valutazione dell'efficacia di questi metodi è necessaria ma difficoltosa, e produce risultati contrastanti (Crisci et al. 2012). L'analisi dei dati genomici è quindi un'area in forte sviluppo, improntata alla ricerca di nuovi approcci capaci di gestire problemi teoricamente più semplici ma di difficile risoluzione pratica, come la velocità di calcolo di statistiche semplici in grandi dataset, o in grado di rispondere a problemi concettualmente più complessi come la valutazione del tasso di falsi postivi o l'identificazione della miglior statistica per identificare segnali di selezione.

In questa tesi ho cercato di rispondere ad alcuni di questi problemi generali in quattro progetti focalizzati su:

A) Il calcolo efficiente di statistiche di genetica di popolazione su dataset estesi usando il calcolo parallelo;

B) L'identificazione di geni selezionati nel fagiolo comune (*Phaseoulus vulgaris*) integrando mediante simulazioni l'informazione della storia demografica della specie;

C) L'identificazione di geni legati al cambiamento delle modalità riproduttiva in *Zootoca vivipara* usando un consenso tra diversi metodi;

D) L'assemblaggio di una regione genomica complessa usando un approccio sperimentale combinato basato su tecnologie di sequenziamento di seconda e di terza generazione.

Introduzione

Le tecnologie di sequenziamento di nuova generazione

La produzione efficiente di sequenze di DNA si può far risalire a Frederick Sanger e collaboratori (Sanger, Nicklen, and Coulson 1977), che svilupparono una tecnologia basata sulla terminazione della catena di DNA. Nei successivi 30 anni, il "Sanger sequencing" diventò la tecnologia più diffusa per il sequenziamento del DNA (Van Dijk et al. 2014). Il risultato più grande raggiunto con questa tecnologia fu il completamento del primo genoma umano nel 2004 (International Human Genome Sequencing Consortium 2004), ottenuto grazie allo sforzo automatizzato e congiunto di molti laboratori.

La produzione del primo genoma umano mostrò però i limiti del metodo Sanger, e fornì la spinta per lo sviluppo di tecnologie di sequenziamento di nuova generazione in grado di fornire una maggiore quantità di informazione con tempi e costi inferiori. La prima tecnologia di seconda generazione, basata sul pirosequenziamento, fu rilasciata nel 2005 dall'azienda 454 Life Sciences (Roche) (Margulies et al. 2005). Nel 2006 fu rilasciata la piattaforma Illumina/Solexa, basata sulla fluorescenza e terminatori reversibili (http://www.illumina.com/technology/next-generation-sequencing/solexa-technology.html), mentre nel 2007 venne presentata la piattaforma SOLiD (Sequencing by Oligo Ligation Detection) dell'azienda Applied Biosystems (Valouev et al. 2008). L'ultima delle principali tecnologie di seconda generazione ad essere introdotta fu la PGM (Personal Genome Machine) della ditta Ion Torrent, basata sull'uso di semiconduttori (Katsnelson 2010). Queste tecnologie sono caratterizzate da tempi, costi e rese diverse (si rimanda a (Glenn 2011) per una sintesi schematica delle performance delle singole tecnologie). La tecnologia più diffusa oggi, in grado di offrire la più elevata resa per singola corsa e il minor costo per base è la tecnologia Illumina (L. Liu et al. 2012), utilizzata anche per la produzione dei dati genomici utilizzati in questa tesi.

Il sequenziamento con le tecnologie di seconda generazione produce numerose letture del genoma (*reads*), *reads* generalmente corte (dalle 75+35 bp della piattaforma SOLiD (ABi) alle 700 bp della piattaforma GS FLX Titanium XL (Roche)) e con un basso tasso di errore (Buermans and den Dunnen 2014). Le

caratteristiche delle *reads* prodotte da queste tecnologie hanno richiesto lo sviluppo di applicazioni bioinformatiche adeguate, come algoritmi di allineamento (Hatem, Bozdağ, and Çatalyürek 2011), algoritmi per l'assemblaggio *de novo* di regioni del genoma (W. Zhang et al. 2011), algoritmi per l'individuazione di polimorfismi (Y. Li et al. 2013) e algoritmi per ridurre i bias introdotti durante la preparazione del campione per il sequenziamento (Hansen, Brenner, and Dudoit 2010).

Un esempio di tecnologia di seconda generazione: la tecnologia Illumina

La tecnologia Illumina, rilasciata dalla ditta Solexa nel 2006, è una tecnologia di sequenziamento basata sul cosiddetto SBS (*sequencing-by-synthesis*, sequenziamento mediante sintesi) (Mardis 2008).

Il DNA da sequenziare viene frammentato e legato ad adattatori specifici, usati per legare i frammenti di DNA a una superficie (flow cell) dove avviene l'amplificazione e il sequenziamento. La frammentazione e il legame degli adattatori può avvenire in un duplice passaggio (frammentazione fisica/enzimatica, seguita da legame con gli adattatori), o in un singolo passaggio (metodologia Nextera (Parkinson et al. 2012), dove un enzima frammenta il DNA e lega gli adattatori nello stesso passaggio). Quest'ultima metodologia, utilizzata in uno dei progetti presentati in questa tesi, è particolarmente indicata in presenza di scarso materiale di partenza.

Il DNA, legato alla superficie di sequenziamento, viene amplificato con la cosiddetta *bridge amplification* (amplificazione a ponte), formando dei cluster di copie di ogni frammento iniziale. I frammenti di ogni cluster vengono denaturati a singoli filamenti, viene legato un primer specifico, e il filamento complementare viene sintetizzato incorporando ciclicamente quattro diversi nucleotidi fluorescenti (con estremità 3'-OH bloccata). In questo modo viene incorporata una singola base alla volta, la cui fluorescenza può essere rilevata con un sensore CCD (*charged coupled device*, dispositivo ad accoppiamento di carica) e convertita in una sequenza di basi. Al termine della lettura, un agente chimico rimuove il fluoroforo dal nucleotide e ripristina la funzionalità della sua estremità 3', permettendo l'incorporazione della base successiva.

I frammenti di DNA possono essere letti da una singola direzione (sequenziamento *single end*) o da entrambe le direzioni dello stesso frammento

(sequenziamento *paired end*). Le due letture dello stesso frammento distano tra loro per una distanza, detta *inner mate distance*, che dipende dalla lunghezza del frammento e delle singole letture, e stimando questa informazione è possibile ottenere un miglior allineamento del frammento in un genoma di riferimento rispetto all'informazione della singola lettura. Ad ogni base delle letture viene inoltre associato un punteggio di qualità, il *Phred Quality Score*, definito come Q = -10 log P, dove P è la probabilità di aver chiamato erroneamente la base (Ewing et al. 1998; Ewing and Green 1998).

Due tra le principali soluzioni commerciali proposte da Illumina, utilizzate in progetti di questa tesi, sono Illumina MiSeq e Illumina HiSeq, che utilizzando una versione molto simile della chimica Illumina. La prima soluzione offre costi inferiori per lo strumento (99.000$ contro i 690.000$ di Illumina HiSeq 2500), un sequenziamento più rapido e *reads* più lunghe, mentre la seconda ha una resa molto superiore e dei costi per MB inferiori. Illumina MiSeq è più indicata per studi di dimensioni contenute, per applicazioni di assemblaggio *de novo* (vista la dimensione delle *reads* più estesa) in piccoli centri di sequenziamento, mentre Illumina Hiseq è indicata per grandi centri di produzione dati, in grado di ammortizzare le spese, e per i progetti dove è richiesta un'alta resa in termini di letture (Glenn 2011).

Le tecnologie di sequenziamento di terza generazione

I limiti delle tecnologie di sequenziamento di seconda generazione portarono allo sviluppo delle cosiddette tecnologie di sequenziamento di terza generazione, caratterizzate da un approccio a singola molecola, sequenziamento in tempo reale e da letture più lunghe (Schadt, Turner, and Kasarskis 2010). Il primo strumento di terza generazione introdotto sul mercato fu il PacBioRS della ditta Pacific Biosciences, basato sull'individuazione dell'attività di una singola DNA polimerasi sul DNA templato. Approcci più recenti prevedono la rilevazione della sequenza della molecola di DNA durante il passaggio di questa in un nanoporo mediante registrazione del segnale elettrico o ottico, o la visualizzazione diretta delle molecole di DNA attraverso tecniche avanzate di microscopia (Schadt, Turner, and Kasarskis 2010). Un esempio recente di sequenziamento basato su nanopori è rappresentata dalla tecnologia Oxford Nanopore, utilizzata in un capitolo di questa tesi.

Ulteriori vantaggi di queste tecnologie, oltre a quelli elencati in precedenza, sono i costi inferiori a parità di basi e l'assenza di amplificazione del DNA in esame, che evita il problema dei duplicati di PCR (ovvero copie dello stesso frammento che vengono letti in cluster diversi e che possono introdurre un distorsione nella genotipizzazione degli individui) e riduce il materiale di partenza necessario. Inoltre il vantaggio di avere lettere più lunghe delle tecnologie di seconda generazione facilita l'assembly *de novo* di regioni complesse e permette l'identificazione diretta degli aplotipi.

Un esempio di tecnologia di terza generazione: la tecnologia Oxford Nanopore

La tecnologia Oxford Nanopore, presentata per la prima volta nel 2012 (Eisenstein 2012), è basata sulla variazione di conduttività elettrica che una molecola di DNA causa durante il passaggio in un nanoporo: il nanoporo, di origine proteica, viene fissato su una superficie polimerica attraversata da una corrente elettrica continua e, quando una molecola di DNA passa al suo interno, viene rilevata la variazione della corrente, variazione caratteristica per ogni tipologia di base. Numerosi nanopori sono presenti sulla superfice del sensore, permettendo l'azione parallela di 512 nanopori, con lettura delle singole molecole in tempo reale senza passare per una fase di sintesi. Le lettura del DNA, ad opera di un sensore, avviene in entrambe le eliche della molecola: un'elica (templato) è infatti unita alla sua elica complementare (complemento) da una forcina molecolare. Il sequenziamento di entrambe le eliche (letture 1D) permette, attraverso un algoritmo interno fornito da un servizio remoto di base calling (Metrichor), di ottenere le cosiddette letture 2D, sequenze di alta qualità formate dall'informazione proveniente da entrambe le eliche. Nonostante questo doppio passaggio, l'accuratezza di questa tecnologia, dichiarata al 95%, è stata stimata da gruppi di ricerca indipendenti attorno a valori più bassi (61,8% - 85%, (Laver et al. 2015; Jain et al. 2015)).

Una particolare commercializzazione di questa tecnologia è il sequenziatore MinION. Questo sequenziatore è caratterizzato da dimensioni compatte (87 g per 10 x 3 x 2 cm), ed è utilizzabile mediante collegamento USB ad un computer. La preparazione del campione per il sequenziamento è veloce (rispetto alle tecnologie di seconda generazione) e l'apparecchio viene fornito con flowcells consumabili, che contengono gli apparati di rilevazione necessari per il sequenziamento. Con un piccolo laboratorio attrezzato è quindi possibile

sequenziare autonomamente i propri campioni, rendendo la piattaforma MinION uno strumento altamente versatile. Il costo di questo strumento è contenuto, il che lo rende un prodotto accessibile a molti centri di ricerca. Lo "starter kit" della fase di beta testing, comprendente il sequenziatore, due flow cells e i reagenti necessari per due esperimenti viene fornito con un pagamento di 1000$ (www.nanoporetech.com). Alcuni svantaggi di questa piattaforma sono legati alla bassa resa per ogni corsa (inferiore sia alla tecnologie di seconda generazione, sia ai diretti concorrenti), che si traduce in un alto costo per Mb sequenziata, e al tasso di errore elevato. Nel corso del mio dottorato ho partecipato alla fase di beta testing di questo strumento, in un progetto volto alla ricostruzione di una regione del sistema maggiore di istocompatibilità nel genoma del camoscio alpino.

Le sotto-rappresentazioni del genoma

Il sequenziamento di interi genomi non è sempre la strategia migliore per rispondere a determinate domande scientifiche, e sono quindi nati diversi approcci per sequenziare sottoinsiemi del genoma sfruttando l'alta resa delle tecnologie di sequenziamento di nuova generazione. Esempi di questi approcci, utilizzati nei progetti presentati in questa tesi, sono il sequenziamento di ampliconi, il RAD-seq e l'RNA-seq (Van Dijk et al. 2014).

La strategia ad ampliconi permette di sequenziare loci specifici del genoma amplificati grazie all'uso di PCR. Durante l'amplificazione, l'amplificato viene legato ad adattatori necessari per il sequenziamento con tecnologie di nuova generazione, permettendone un successivo sequenziamento. I vantaggi di questo approccio sono legati alla possibilità di sequenziare brevi porzioni del genoma con un'elevata copertura.

La metodologia RAD-seq permette di sequenziare un campione casuale del genoma, sfruttando la frammentazione del genoma ad opera di enzimi di restrizione (Baird et al. 2008b). Una piccola parte del genoma, quella associata ai siti di taglio dell'enzima utilizzato, viene sequenziata con tecnologie di nuova generazione, rendendo questa tecnologia ideale per genotipizzare molti individui in specie con genomi molto estesi. Questa tecnologia può essere applicata sia ad organismi modello (per i quali esiste un genoma di riferimento su cui mappare le *reads*), sia ad organismi non modello, andando a creare un catalogo di loci *de*

novo aggregando le *reads* identiche in sequenze uniche. Questa metodologia permette quindi di individuare polimorfismi su un ampio spettro di organismi, rendendola uno strumento rilevante per gli studi di genetica di popolazioni (Davey et al. 2010).

Un'altra strategia è legata al sequenziamento del trascrittoma, cioè l'intero set di trascritti di una cellula, tessuto o intero organismo in un particolare momento o stadio di sviluppo o in talune condizioni fisiologiche. Per RNA-seq si intende il sequenziamento con tecnologie di nuova generazione dell'intero trascrittoma, con una serie di passaggi *ad hoc* per il tipo di molecola in esame (ad esempio, enzimi per evitare la degradazione dell'RNA e l'uso di trascrittasi inverse). La metodologia RNA-seq non richiede l'uso di un genoma di riferimento ed è quindi adatta anche quando si studiano organismi non modello (Wang, Gerstein, and Snyder 2009)

Inferenze evolutive su dati genomici

L'aumento della quantità di informazione disponibile ha introdotto nuove sfide nell'analisi dei dati genetici, sia da un punto di vista teorico che pratico. Infatti, il nuovo tipo di dato ha aumentato sia la quantità che il tipo di informazione disponibile (si pensi ad esempio alla possibilità di individuare lunghi aplotipi nel genoma), portando a una rivalutazione dei metodi ideati per l'analisi su singoli marcatori. Un'altra sfida è legata alle risorse computazionali necessarie per l'analisi di questa elevata quantità di informazione: i vecchi metodi sono stati infatti sviluppati per dataset limitati, e in alcuni casi non possono essere applicati quando il dataset supera una certa dimensione. Vi è quindi la necessità, nell'analisi dei dati genomici, di sviluppare nuovi metodi o rendere vecchie metodologie compatibili con il nuovo tipo di dato. Questo problema è stato affrontato nel primo capitolo della tesi, dove ho presentato 4P, un software in grado di calcolare con efficienza e velocità semplici statistiche di genetica di popolazione su pannelli di polimorfismi di provenienza genomica.

In questa parte dell'introduzione, cercherò di descrivere i principali approcci all'analisi dei dati genomici, soffermandomi su una delle due principali aree di interesse della genetica di popolazioni: la ricerca delle determinanti geniche della selezione naturale o artificiale. Questa parte è stata approfondita soprattutto nel secondo e nel terzo capitolo di questa tesi, che trattano rispettivamente

l'identificazione di geni potenzialmente selezionati durante il processo di domesticazione del fagiolo comune in Mesoamerica e l'associazione di geni con il cambiamento della modalità riproduttiva in *Zootoca vivipara*.

La ricerca di loci soggetti al processo di selezione

I segnali di selezione a livello del genoma vengono generalmente individuati comparando il pattern di variazione genetica del sito/regione analizzati con la distribuzione della variabilità genomica, per limitare l'effetto confondente dei processi demografici. Le principali tracce di selezione si basano su: l'alta proporzione di mutazioni che alterano la funzione di un gene (specie specifiche), la riduzione della diversità genetica, la presenza di alleli derivati ad alta frequenza, il differenziamento tra popolazioni e la presenza di lunghi aplotipi (P C Sabeti et al. 2006).

Aumento di mutazioni che alterano la funzione di un gene

La selezione positiva può aumentare il tasso di fissazione di una mutazione che comporta un cambiamento benefico nella funzionalità di un gene (WH Li, Wu, and Luo 1985), cambiamento che può essere rilevato confrontando le sequenze di DNA di specie diverse. Questo tipo di selezione può essere identificata su un'ampia scala temporale, ma richiede una serie di cambiamenti selezionati per essere distinta dal tasso di mutazione neutrale del background genomico. Questo segnale può essere identificato confrontando il tasso delle sostituzioni non sinonime per sito non sinonimo con il tasso di sostituzioni sinonime per sito sinonimo o altri cambiamenti neutrali, confrontando il tasso con altre linee o con la diversità interna alla specie. I due principali test che utilizzando questo approccio sono Ka/Ks (WH Li, Wu, and Luo 1985) e il test di McDonald-Kreitman (McDonald and Kreitman 1991).

Questi due test possono essere svolti su un qualsiasi set di regioni codificanti, quindi possono essere usate in presenza di dati genomici (a condizione che le regioni geniche siano annotate in un genoma di riferimento vicino), in dati RAD-seq (a patto che siano coperte delle regioni geniche e ci sia un genoma di riferimento annotato vicino), in sequenze geniche amplificate con ampliconi e soprattutto con un approccio di tipo RNA-seq.

Riduzione della diversità genetica

Quando un allele benefico aumenta la sua frequenza nella popolazione, anche le varianti in posizioni vicine dello stesso cromosoma aumentano di frequenza. Questo effetto, detto effetto *hitchhiking*, altera il pattern di diversità genetica della regione, azzerando la diversità vicino all'allele selezionato e riducendola in una finestra più ampia (Maynard Smith and Haigh 1974). Le nuove mutazioni che appaiono nella regione sono inizialmente a bassa frequenza, creando quindi un segnale di regione a bassa diversità con un eccesso di mutazioni rare. La dimensione della regione genomica affetta da questa riduzione della diversità dipende dalla forza della selezione positiva: più la fissazione è rapida, più la regione interessata è ampia. L'identificazione di questa regione è semplice (seppure spesso mimata da effetti demografici), mentre è difficoltoso stabilire quale variante causale sia stata selezionata (Oleksyk, Smith, and O'Brien 2010). I test statistici usati per identificare questa traccia di selezione sono test basati su semplici statistiche di diversità genetica, il test di Hudson-Kreitman-Aguadè (Hudson, Kreitman, and Aguadé 1987) e statistiche in grado di catturare variazioni nella distribuzione delle frequenze alleliche (ad esempio la D di Tajima (Tajima 1989)).

Gli approcci basati su questo segnale calcolano la diversità genetica, usando stimatori più o meno complessi, in numerosi loci o finestre del genoma. I valori stimati per ogni unità in esame vengono confrontati per individuare loci/finestre del genoma con bassa variabilità rispetto agli altri, che contengono putativamente delle varianti selezionate positivamente. Questo tipo di segnale può essere rilevato in qualsiasi tipo di dato genomico, ma risente molto dell'impatto confondente della storia demografica della specie.

Alta frequenza di alleli derivati

Il processo mutazionale introduce nuovi alleli in una popolazione, con frequenze alleliche più basse rispetto agli alleli ancestrali (G. A. Watterson and Guess 1977). In presenza di selezione, gli alleli derivati associati ad alleli benefici possono aumentare la loro frequenza, creando delle regioni che contengono molti alleli derivati ad alta frequenza. Questo tipo di segnale può essere individuato quando l'allele ancestrale è noto (cioè è inferito dall'allele presente in una specie vicina,

assumendo che la mutazione si avvenuta dopo la divergenza tra le specie). Non è influenzato da eventi demografici, ma può essere influenzato da popolazioni suddivise (Przeworski 2002). Le statistiche usate per testare questo segnale sono la H di Fay e Wu (Fay and Wu 2000) e la F di Fu e Li. (Fu and Li 1993).

Questo segnale può essere testato su qualsiasi tipo di dato genomico, ma occorre prima determinare quale degli alleli di un polimorfismo rappresenti lo stato ancestrale. In specie modello, come l'uomo, questa informazione può essere ricavata facilmente, in quanto sono note molte varianti omologhe in specie vicine. In organismi non modello questa informazione è di difficile valutazione, rendendo l'utilizzo di questo approccio sconsigliato in tali sistemi.

Differenziamento tra popolazioni

Quando popolazioni geograficamente distinte sono sottoposte a pressioni selettive diverse, la selezione modifica le frequenze di alcuni alleli in una popolazione ma non nell'altra. Differenze nelle frequenze alleliche tra popolazioni possono quindi essere un segnale di selezione positiva. Questo tipo di segnale può essere identificato quando le popolazioni sono almeno parzialmente isolate da un punto di vista riproduttivo. La statistica principale usata per testare queste differenze è l'F_{st}.

L'F_{st} è stato definito da Wright (S. Wright 1949) come la parte di variabilità genetica che può essere attribuita alle differenze tra demi, e può essere stimato con diversi stimatori (ad esempio (Weir and Cockerham 1984; Nei 1973)). Valori outlier di F_{st} possono essere identificati in modi differenti. Un primo approccio generale prevede la generazione di una distribuzione attesa di F_{st} in neutralità date delle frequenze alleliche iniziali o un modello demografico: loci con valori di F_{st} outlier rispetto a questa distribuzione attesa in neutralità rappresentano loci putativamente selezionati (Foll and Gaggiotti 2008). Un altro approccio prevede il calcolo di F_{st} in molte posizioni del genoma, identificando le regioni con i valori più estremi di F_{st} come regioni selezionate (Akey et al. 2002). Questi approcci permettono di identificare eventi selettivi con diverse intensità, ed sono influenzato dalle frequenze alleliche pre-selezione, dalla migrazione e dalla deriva genetica. Un limite dell'F_{st} è legato al fatto che non è possibile stabilire con certezza quale delle due popolazioni confrontate sia stata sottoposta a selezione. Per ovviare a questo problema è stata sviluppata una derivazione dell'F_{st}, la statistica *"locus*

specific branch length" (LSBL) (Shriver et al. 2004). Questa statistica confronta i valori di F_{st} di tre o più popolazioni diverse per identificare quale di queste sia il reale bersaglio della selezione.

Questa statistica può essere calcolata su qualsiasi tipologia di dato genomico, ma può essere influenzata dalla storia demografica delle popolazioni.

Lunghi aplotipi

Quando un allele è selezionato, la sua frequenza può aumentare così velocemente che la ricombinazione non riesce a rompere l'associazione fisica con gli alleli vicini. La selezione crea quindi una traccia genomica non attesa in neutralità: un allele ad alta frequenza con un'associazione a lungo raggio con altri alleli. Approcci basati su questo segnale permettono di identificare eventi selettivi non ancora terminati, ma solo per eventi relativamente recenti (0.1 Ne generazioni, (Pfaffelhuber, Lehnert, and Stephan 2008)). Questo segnale può essere identificato confrontando la diversità aplotipica o con dei test derivati dall'EHH (*extended haplotype homozygosity*) (Pardis C. Sabeti et al. 2002).

Questi approcci richiedono la presenza di lunghi tratti del genoma per identificare l'associazione a lungo raggio tra alleli, pertanto il loro uso principale è su dataset di genomi completi in organismi in cui la variazione del tasso di ricombinazione e la fase dei dati genetici sono noti (prevalentemente organismi modello).

4P: Parallel Processing of Polymorphism Panels

Uno dei principali problemi nell'analisi di dati genomici, conseguenza della grande quantità di informazione prodotta, è legato alla gestione di questa informazione e alla sua elaborazione in tempi ragionevoli. In genetica di popolazioni questo problema è amplificato, in quanto l'analisi è spesso rivolta a un campione numeroso, formato da più di un individuo. All'inizio del mio percorso di dottorato il calcolo in parallelo di statistiche intra-popolazione ed inter-popolazioni usate in genetica di popolazioni su grandi pannelli di polimorfismi non era ancora disponibile, rendendo l'analisi di questi dati un'operazione lenta e laboriosa. Per affrontare questo problema, ho contribuito alla scrittura del codice sorgente di 4P, un software in grado di calcolare alcune statistiche comunemente usate in genetica di popolazioni in modo rapido ed efficiente. 4P, supportando i formati comunemente utilizzati negli studi di genetica di popolazioni, permette di analizzare dataset di milioni di SNPs tipizzati in diversi individui provenienti da diverse popolazioni. Le performance di 4P sono state valutate usando pannelli di SNPs provenienti dal genoma umano e da simulazioni, rivelando come 4P sia più veloce di altri software o pacchetti che permettono di calcolare le stesse statistiche. La portabilità e i bassi requisiti di sistema di 4P rendono il software ideale in analisi esplorative su grandi pannelli di dati genomici o in studi di simulazione con dataset multipli. I risultati di questo lavoro sono stati pubblicati sulla rivista Ecology and Evolution (Benazzo, Panziera, and Bertorelle 2015).

Introduzione

Le moderne tecnologie di sequenziamento hanno aumentato drasticamente la capacità di individuare polimorfismi nel campione analizzato, ed è oggi possibile disporre per alcuni organismi di dataset di milioni di polimorfismi in migliaia di individui. Un esempio di dataset di queste dimensioni è il set di polimorfismi a singolo nucleotide caratterizzati nell'uomo nel progetto 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015), che nella fase 3 del progetto ammontano a 84.7 milioni per 2504 individui provenienti da 26 popolazioni. Questi dataset possono essere prodotti con diverse tecnologie, più o meno adatte all'utilizzo con organismi modello o non modello. Ad esempio, una

caratterizzazione della variabilità genomica in organismi non modello può essere ottenuta grazie alla tecnologia RAD-seq (Davey et al. 2011) che, abbinata alla pipeline implementata nel software STACKS (Catchen et al. 2013), permette di individuare polimorfismi in posizioni casuali nel genoma in un set di individui anche in assenza di un genoma di riferimento. Il dataset di polimorfismi ottenuto può essere analizzato per calcolare statistiche, come il differenziamento delle popolazioni e il tasso di admixture, che sono essenziali per l'identificazione di regioni del genoma responsabili di adattamento e speciazione e permettono di affrontare problemi classici nella genetica di popolazioni da una nuova prospettiva. (ad esempio (Keller et al. 2013; Hohenlohe et al. 2013)). Il trend attuale è quindi quello di dataset di polimorfismi sempre più grandi per numerosi organismi, dataset la cui analisi rappresenta una sfida sia a livello teorico, sia a livello pratico.

Le statistiche in grado di stimare i livelli di variazione genetica intra-popolazione ed inter-popolazioni possono essere utilizzate sia in analisi esplorative su dati reali, sia per riassumere la diversità genetica in un grande numero di dataset simulati (dataset con dimensioni paragonabili a quelli ottenuti da dati reali). Quando queste statistiche sono calcolate su milioni di loci i tempi computazionali richiesti per un'analisi con i software disponibili sono lunghi, a volte proibitivi. Ad esempio, STACKS e PLINK (Catchen et al. 2013; Purcell et al. 2007), due software sviluppati rispettivamente per l'analisi di dati RAD-seq e per studi di associazione, utilizzano algoritmi di calcolo di tipo seriale che rallentano i tempi computazionali, mentre Adegenet e PopGenome (Jombart and Ahmed 2011; Pfeifer et al. 2014), due pacchetti R per l'analisi di dataset genomici di SNPs, hanno solo poche funzioni ottimizzate per il calcolo parallelo. Queste limitazioni obbligano i ricercatori a pensare a soluzioni alternative per velocizzare l'analisi, come progettare scripts/programmi ad hoc o partizionare il dataset prima dell'analisi. In questo modo un processo semplice come il calcolo di una statistica in un dataset diventa un processo laborioso e soggetto a numerose fonti di errori, che nel complesso rallentano l'attività di ricerca.

Per ovviare a questi problemi è stato sviluppato 4P (Parallel Processing of Polymorphism Panels), un software scritto in ANSI C che utilizza le librerie di calcolo parallelo OpenMP. 4P è stato specificamente pensato per il calcolo rapido di statistiche di genetica di popolazioni da pannelli di SNPs, e la sua

implementazione in parallelo permette di aumentare la velocità di calcolo in sistemi a più processori come PC desktop e server.

Caratteristiche del software

Formati supportati

Un pannello di SNPs può essere descritto come una matrice N x M, dove N rappresenta il numero di individui aploidi (o cromosomi omologhi negli organismi diploidi) ed M il numero di SNPs analizzati. Questo tipo di struttura bidimensionale è ottimale per la manipolazione utilizzando il linguaggio C e la parallelizzazione, in quanto il calcolo delle statistiche può essere facilmente suddiviso tra i vari core del sistema. Nello specifico, un certo numero di colonne/SNPs viene assegnato ad uno specifico core grazie all'API OpenMP, che permette di calcolare contemporaneamente su più SNPs le statistiche desiderate andando a partizionare l'informazione. Questa operazione è possibile in quanto le statistiche implementate in 4P non utilizzano l'informazione di tutti i diversi loci, ma l'informazione di ogni locus rappresenta una unità indipendente da quella degli altri marcatori. Il dataset dei polimorfismi viene immagazzinato da 4P nella memoria RAM del sistema prima dell'elaborazione, utilizzando delle routine ottimizzate per ridurre i tempi di lettura dell'informazione da disco rigido. Il fattore limitante per la quantità di informazione analizzabile (la dimensione della matrice N x M) è quindi legato alla quantità di memoria RAM disponibile nel sistema.

I formati attualmente supportati da 4P sono:

- Plink
- Arlequin
- Variant Call Format

Alcuni di questi formati non contengono tutta l'informazione necessaria a 4P per calcolare alcune statistiche. Il software richiede quindi ulteriori file di input, nello specifico un file in cui viene specificata la popolazione di origine di ogni individuo quando si utilizza il formato Variant Call, e un file contenente l'informazione dell'allele ancestrale di una posizione polimorfica con i formati Variant Call e Plink.

Statistiche calcolate

4P permette di calcolare numerose statistiche usate in genetica di popolazioni per ogni locus, permettendo anche il calcolo della media e della varianza tra i diversi loci.

Il numero di alleli per ogni posizione supportato da 4P varia da 1 (sito monomorfico) a 4.

Le statistiche calcolate sono:

- Frequenze alleliche
- Eterozigosi attesa e osservata
- Spettro delle frequenze alleliche (folded e unfolded, a una o più dimensioni)
- G_{st} (3 formulazioni) (Nei 1973; Nei and Chesser 1983; Hedrick 2005)
- D di Jost (Jost 2008)
- F_{st} (Weir and Cockerham 1984)
- Proporzione degli alleli condivisi tra coppie di individui.

Esempi di applicazioni

4P, per la sua natura di programma semplice, può essere usato per calcolare tutte le statistiche implementate in presenza di una quantità molto grande di dati, come ad esempio nel caso di dati genomici. Un esempio di applicazione è stata fatta nel lavoro di (Tassi et al. 2015), dove 4P viene utilizzato per calcolare l'F_{st} in un dataset di decine di migliaia di SNPs in 1130 individui provenienti da 24 popolazioni.

Oltre a questa applicazione, più tradizionale, 4P può essere utilizzato all'interno di pipeline più complesse, come ad esempio in una analisi di *Approximate Bayesian Computation*. In questo tipo di applicazione 4P può essere impiegato per calcolare le statistiche riassuntive nel dataset osservato e nei dataset generati da simulazioni in coalescente.

Performance e confronto con altre applicazioni

Le performance di 4P sono state valutate su un sistema Intel Xeon X5650 a 16 core a 2.66 GHz, con 32 Gb di RAM. Il file utilizzato per valutare le performance è il prodotto di una simulazione con *fastsimcoal* (Laurent Excoffier and Foll 2011; Laurent Excoffier et al. 2013), in cui sono stati simulate due popolazioni di 500 individui (diploidi) che si sono separate tra loro 1000 generazioni fa. Sono state prodotte diverse versioni di questo file (in formato Arlequin), con un numero variabile di polimorfismi.

Nelle seguenti tabelle sono riportati rispettivamente i tempi computazionali per il calcolo dell'eterozigosi osservata e attesa (Tabella 1) e di cinque diverse misure di differenziamento (F_{st}, tre formulazioni di G_{st} e D di Jost) (Tabella 2) al variare del numero di polimorfismi e del numero di core impiegati. Questi valori includono sia il tempo di caricamento del file, sia il tempo di elaborazione.

SNPs	1 core	2 cores	4 cores	8 cores	16 cores
1000	0,62	0,53	0,52	0,55	0,48
10000	2,86	1,71	1,28	1,02	0,95
100000	27,30	15,97	9,41	6,29	5,76
1000000	192,82	112,54	69,22	49,60	42,30

Tabella 1. Tempi di calcolo (in secondi) dell'eterozigosi attesa e osservata al variare del numero di SNPs e di processori utilizzati.

SNPs	1 core	2 cores	4 cores	8 cores	16 cores
1000	0,61	0,55	0,53	0,48	0,51
10000	1,86	0,97	0,91	0,85	0,82
100000	12,05	5,89	3,74	2,88	2,54
1000000	87,81	54,88	40,81	31,83	29,33

Tabella 2. Tempi di calcolo (in secondi) di cinque statistiche di differenziamento al variare del numero di SNPs e di processori utilizzati.

Il grafico seguente (Figura 1), tratto dall'articolo pubblicato (Benazzo, Panziera, and Bertorelle 2015), permette di valutare qualitativamente la variazione dei tempi computazionali al variare del numero dei processori.

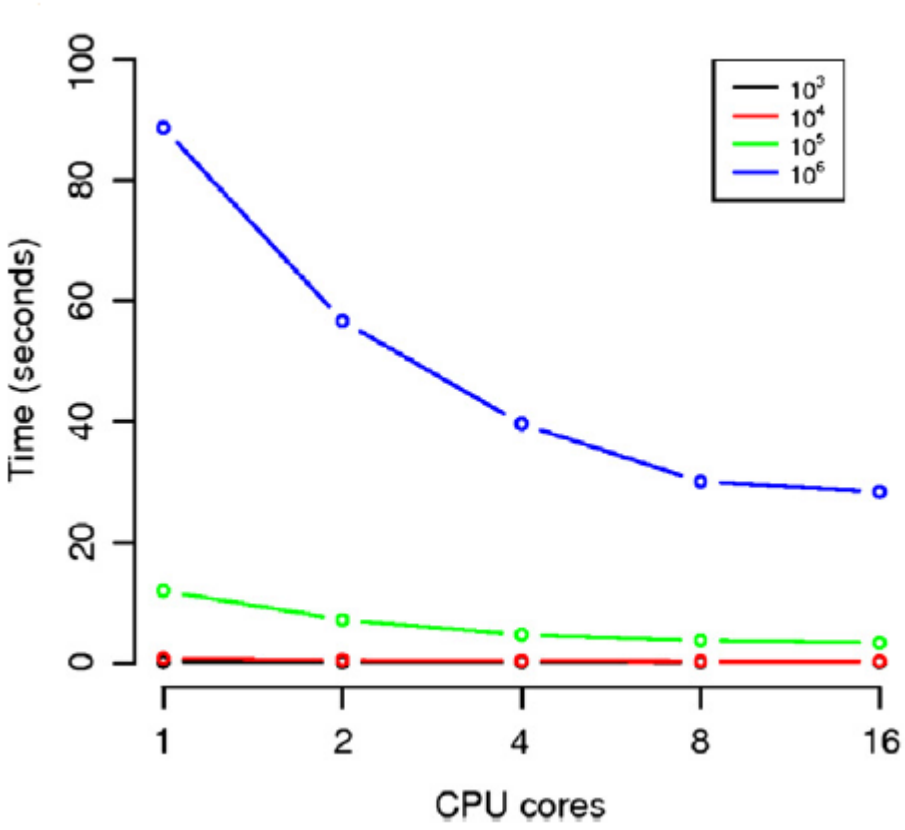

Figura 1. Tempi di calcolo (in secondi) di cinque statistiche di differenziamento al variare del numero di cores e di polimorfismi utilizzati (colori diversi rappresentano dataset con un diverso numero di SNPs).

Dal grafico si può notare come con un dataset di piccole dimensioni (ad es. 1000 SNPs) l'utilizzo di più core non aumenti significativamente la velocità dell'analisi, in quanto il dataset è piccolo e il calcolo è quasi istantaneo. Con l'aumentare del numero di dati l'impatto della parallelizzazione è più rilevante. Questo risultato è importante se si considerano i campi di utilizzo di 4P: in presenza di analisi ripetute, infatti, differenze di qualche secondo nella singole ripetizioni diventano via via più rilevanti all'aumentare del numero di ripetizioni.

I risultati mostrano come, aumentando il numero di core utilizzati, il tempo dell'analisi raggiunge rapidamente un punto dove le variazioni sono minime. Questo risultato, unito al fatto che durante l'analisi di questi dataset la quantità di

memoria RAM utilizzata non ha mai superato i 2Gb, suggerisce come utilizzando 4P la maggior parte dei computer di fascia medio-bassa (con pochi core e memoria RAM) è in grado di gestire il calcolo di statistiche utilizzando pannelli di SNPs di grandi dimensioni.

Confronto con PLINK

PLINK è un software che permette di svolgere numerose analisi su dati genomici. Esempi di questi analisi sono test di associazione, identificazione di stratificazione di popolazione e calcolo di statistiche riassuntive per il controllo della qualità, tra cui il calcolo dell'eterozigosi attesa e osservata.

Il confronto con PLINK è stato fatto calcolando i due tipi di eterozigosi in due dataset.

Il primo dataset è il dataset di dati simulati con fastsimcoal descritto nel paragrafo precedente. PLINK, non permettendo il calcolo in parallelo, è stato testato unicamente su singolo processore, mentre 4P con tre diverse quantità di processori. Il risultato del confronto è riportato nella seguente tabella:

SNPs	PLINK (1 core)	4P (1 core)	4P (4 cores)	4P (16 cores)
1000	0,92	0,62	0,52	0,48
10000	4,52	2,86	1,28	0,95
100000	46,88	27,30	9,41	5,76
1000000	474,83	192,82	69,22	42,3

Tabella 3. Tempi di calcolo (in secondi) dell'eterozigosi al variare del numero di SNPs, utilizzando PLINK e 4P con un diverso numero di processori.

I risultati sono rappresentati nel seguente grafico (Figura 2), tratto dall'articolo pubblicato (Benazzo, Panziera, and Bertorelle 2015).

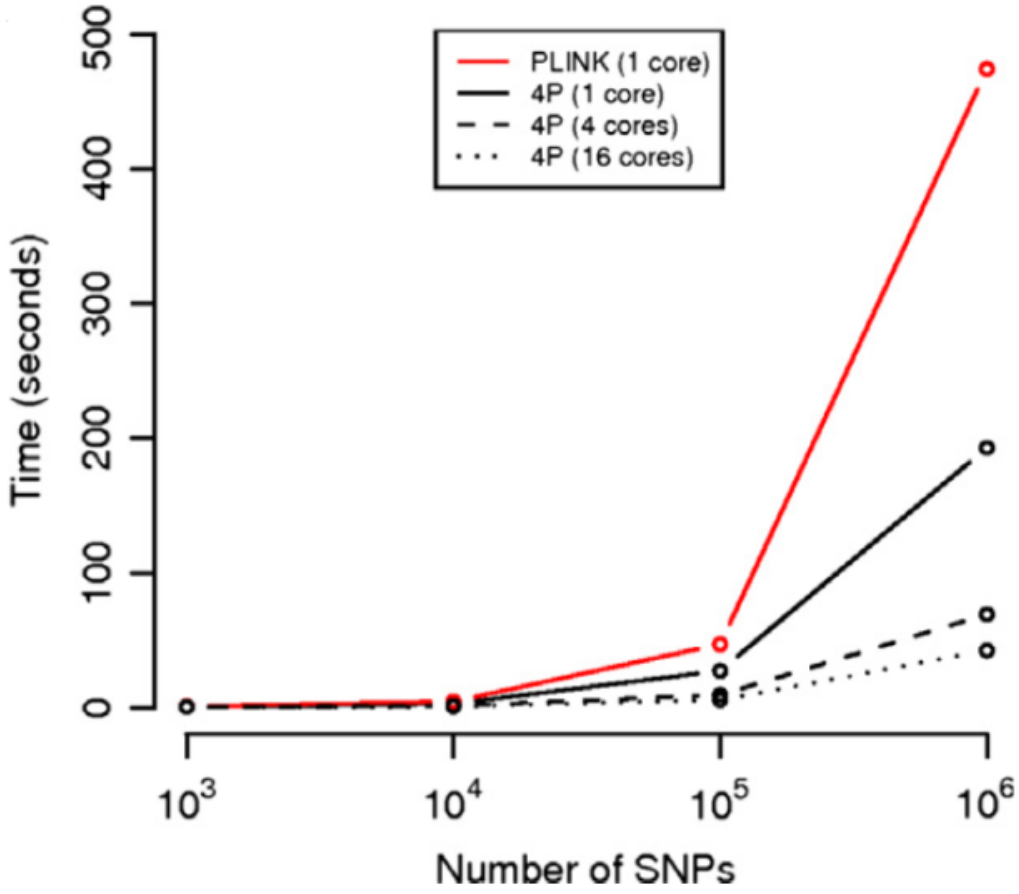

Figura 2. Tempi di calcolo (in secondi) dell'eterozigosi in funzione del numero di polimorfismi del dataset e del tipo di software utilizzato.

I risultati mostrano come 4P, utilizzando un solo processore, risulti essere più veloce di PLINK da 1,5 a 2,5 volte, in base al numero di SNPs analizzati. Utilizzando 4P con più processori questa differenza è amplificata in maniera sostanziale. Ad esempio, utilizzando il dataset con un milione di polimorfismi, 4P con 16 processori risulta più veloce di 11 volte di PLINK.

Il secondo dataset testato è un dataset di dati reali provenienti dalla Fase 1 del 1000 Genome Project ((The 1000 Genomes Project Consortium 2012), disponibile su ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) Questo dataset, in formato PLINK, è formato da 1092 individui e circa 3 milioni di polimorfismi a singolo nucleotide. Anche in questo caso sono stati calcolati i due tipi di eterozigosi. Il tempo computazionale utilizzando PLINK è pari a 29 minuti e 45 secondi, mentre utilizzando 4P con 16 processori è pari a 9 minuti e 9 secondo, mostrando una riduzione di circa 3 volte. I tempi computazionali sono più elevati

rispetto a quelli dell'analisi precedenti in quanto il processo di caricamento di un file ped rispetto a un file arp è più lento. Considerando solamente i tempi necessari per il calcolo delle statistiche, la differenza in termini di tempo è analogo a quello osservato nell'analisi di dataset simulati (circa 11 volte).

Confronto con PopGenome

PopGenome (Pfeifer et al. 2014) è un pacchetto di R (R Core Team 2015) sviluppato per la gestione dei dati genomici, permettendo sia il calcolo di numerose statistiche di genetica di popolazioni, sia la loro integrazione in algoritmi più complessi.

Il confronto fra 4P e PopGenome è stato fatto calcolando indici di variabilità genetica in un milione di SNPs provenienti dal cromosoma 1, tipizzati nei 1092 individui della Fase 1 del 1000 Genome Project. Nel confronto è stata valutata sia la velocità di lettura del file, sia la velocità di calcolo delle statistiche.

Il confronto sulla lettura del file è stato fatto utilizzando la funzione read-VCF di PopGenome, ottimizzata per il tipo di dato, e 4P. Il tempo di lettura è risultato pari a circa 11 minuti utilizzando PopGenome, 5 minuti utilizzando 4P, mostrando come 4P riesca a dimezzare i tempi di lettura rispetto a PopGenome.

L'analisi delle performance dei due metodi è stata fatta calcolando diversi indici di diversità in una macchina più piccola rispetto a quella utilizzata in precedenza (con 8 cores e 8 Gb di RAM). A causa dell'enorme utilizzo di memoria richiesto da PopGenome, non è stato possibile analizzare il dataset in un singolo passaggio, ma abbiamo dovuto suddividere il cromosoma in intervalli di 1000 paia basi. In ognuno di questi intervalli è stata calcolata la diversità nucleotidica utilizzando la funzione diversity.stat, impiegando in totale circa 10 minuti e 8 Gb di RAM. Nello stesso dataset non partizionato sono state calcolate l'eterozigosi osservata e attesa e lo spettro delle frequenze alleliche (un numero maggiore di statistiche) utilizzando 4P. I tempi computazionali variano da 2 minuti e 30 secondi utilizzando un singolo core a circa 30 secondi utilizzando 8 cores, senza mai consumare più di 3 Gb di memoria RAM.

Conclusione

In questo capitolo della tesi ho presentato 4P, un programma in C per l'analisi in parallelo di pannelli di SNPs con elevata dimensione. 4P permette il calcolo di diverse statistiche di genetica di popolazioni, e può essere utilizzato sia per dataset reali che simulati. Il confronto con altri software mostra come 4P sia più veloce di software o pacchetti di R paragonabili, non richieda script *ad hoc* e possa essere usato facilmente in computer e server di fascia medio-bassa.

Dove trovare 4P

Il manuale, il codice sorgente, il software (per Unix, MacOs e Windows) e alcuni file di esempio possono essere scaricati al seguente indirizzo https://github.com/anbena/4p. Il software è rilasciato con licenza GPLv3.

Glossario

Core e parallelizzazione

In informatica per core si definisce il nucleo fisico dell'unità di elaborazione centrale (CPU) di un computer, deputata all'esecuzione delle istruzioni di un programma presente nella memoria RAM. Nel 2005 i sistemi a singolo core hanno raggiunto il loro valore limite nelle frequenze operative, ed è iniziata la commercializzazione di sistemi multi-core, dove la CPU è costituita da diversi core (in genere 2, 4, 8 o 16) in comunicazione tra loro. Questo ha portato allo sviluppo di tecniche di parallelizzazione, permettendo l'esecuzione simultanea del codice di un programma su più core dello stessa CPU.

API

Acronimo di Application Programming Interface, rappresenta un set di strumenti specifici per il programmatore per svolgere un determinate compito all'interno di un programma.

OpenMP

OpenMP è un'API per la creazione di applicazioni parallele su sistemi a memoria condivisa.

Formato Arlequin

Il formato Arlequin, usato per l'analisi con l'omonimo software (Laurent Excoffier and Lischer 2010), è supportato da 4P solo quando rappresenta un dataset di dati simulati prodotti con *fastimcoal* (Laurent Excoffier and Foll 2011; Laurent Excoffier et al. 2013), in formato arp. In questo file sono immagazzinate tutte le informazioni necessarie, dalle popolazioni allo stato allelico del polimorfismo (ancestrale/derivato). Nel caso si utilizzino dati reali, è necessario convertire il file arp in uno degli altri due formati supportati.

Formato PLINK

Questo formato, implementato nell'omonimo software (Purcell et al. 2007), si compone di due file: un file ped, in cui sono immagazzinate informazioni sul genotipo, la popolazione di origine e il fenotipo del campione, e un file map, contenente informazioni sulla posizione dei polimorfismi nel genoma dell'individuo. Questo formato presenta anche un'estensione in formato binario (file bed), la cui lettura non è supportata da 4P ma che può essere facilmente ottenuto da un file bed utilizzando PLINK.

Variant Call Format

Questo tipo formato si compone di un singolo file vcf contenente numerose informazioni sul genotipo del campione. Questo formato rappresenta oggi uno standard nei dataset di polimorfismi ed è presente in numerosi database genomici, ad esempio quello del progetto 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015)

Determinanti genetiche e domesticazione: il caso del fagiolo comune (*Phaseolus vulgaris*).

La ricerca di geni o tratti del genoma sotto selezione rappresenta uno dei principali campi di studio della genetica di popolazioni. Gli effetti della selezione naturale sono spesso indistinguibili da quelli di eventi demografici, e lo studio di metodi per distinguere tra questi due fenomeni è tutt'ora un'area di forte interesse. Nel caso della domesticazione questo problema è amplificato, in quanto forti pressioni selettive per particolari tratti sono associate ad eventi demografici drastici come forti riduzioni nella dimensione della popolazione.

In questo capitolo della mia tesi affronterò questo problema parlando di uno studio a cui ho partecipato inerente l'effetto della domesticazione nel fagiolo comune (*Phaseolus vulgaris*), una specie autogama caratterizzata da eventi di domesticazione multipli nel corso della sua storia evolutiva. Questo studio è stato svolto in collaborazione con il gruppo di Roberto Papa dell'Università Politecnica delle Marche, il gruppo di Massimo Delledonne delll'Università di Verona e il gruppo di Zoran Nikoloski del Max Planck Institute of Molecular Plant Physiology di Potsdam (Germania), e comprende una parte di genetica di popolazioni (alla quale ho contribuito) e una parte di analisi di espressione genica e di studio di network genici.

Lo studio si è focalizzato sul processo di domesticazione del fagiolo comune in Mesoamerica, utilizzando un campione di genotipi di fagiolo comune mesoamericani domesticati e selvatici. Insieme a questi sono stati utilizzati anche alcuni genotipi andini come riferimento. All'epoca dell'inizio di questo studio non era ancora disponibile un genoma di riferimento evolutivamente vicino agli individui mesoamericani, ed è stato quindi assemblato un trascrittoma *de novo* di 27.243 contigs utilizzando un gruppo rappresentativo di genotipi mesoamericani e andini. Il trascrittoma è stato utilizzato come riferimento per mappare le *reads* di 21 individui, ottenendo un dataset di circa 190,000 polimorfismi a singolo nucleotide (SNPs).

Nella popolazione mesoamericana domesticata si sono osservate una forte riduzione nella diversità nucleotidica (circa 60%) e la fissazione di alleli polimorfici

nella forma selvatica. Per discriminare gli effetti della selezione da quelli demografici, abbiamo utilizzato le informazioni demografiche disponibili per la specie in esame per ricavare la distribuzione nulla (assumendo assenza di selezione naturale) di alcune statistiche utili per individuare loci sotto selezione positiva. I loci con valori outliers in questa distribuzione, il 9% dei totali, sono stati classificati come putativamente sotto selezione. Sono stati analizzati inoltre i livelli di espressione dei geni del trascrittoma, osservando una riduzione nella diversità dell'espressione genica nella popolazione domesticata, riduzione più ampia nei geni sotto selezione. L'analisi dei network ha confermato questi pattern, mostrando strutture di comunità distinte tra le popolazioni domesticate e selvatiche, con un arricchimento per differenti funzioni molecolari. I risultati di questo lavoro sono stati pubblicati sulla rivista The Plant Cell. (Bellucci et al. 2014).

Questo studio, il tema principale di questo capitolo, è stato ulteriormente ampliato studiando un dataset di 43 geni sequenziati disponibili in letteratura. Dopo una caratterizzazione della diversità genetica a livello di questi loci sono stati individuati i geni putativamente sotto selezione durante la domesticazione del fagiolo comune nel Mesoamerica utilizzando una versione aggiornata e *ad hoc* del metodo utilizzato nello studio precedente. I risultati di questo studio confermano il trend generale di riduzione della diversità genetica nella popolazione domesticata e forniscono evidenze di selezione a livello di alcuni geni. Questi sono stati confrontati con i risultati dello studio sul trascrittoma e con quelli di uno studio recente che utilizza le informazioni dell'intero genoma di fagiolo comune (Schmutz et al. 2014), evidenziando una parziale sovrapposizione dei risultati ed alcuni limiti dei metodi utilizzati.

L'identificazione di geni sotto selezione nel fagiolo comune mesoamericano con dati RNA-seq

Introduzione

La domesticazione delle piante è un campo di studi da tempo oggetto di forte interesse nell'ambito della biologia evoluzionistica. Come affermato dallo stesso Charles Darwin, la domesticazione può essere considerata un grande esperimento evolutivo (Darwin 1868), mentre dal punto di vista dello sfruttamento delle piante a fini agricoli, lo studio della domesticazione è la chiave per lo sviluppo di strategie di procreazione e per l'identificazione di varianti genetiche utili.

Il processo di domesticazione ha plasmato numerosi tratti nel corso dei millenni, tratti che permettono oggi di distinguere le specie domesticate dalle loro forme selvatiche. Questi tratti, comuni a molte specie domesticate, contribuiscono alla cosiddetta "sindrome della domesticazione" (Paul Gepts and Papa 2002), e includono la dimensione, la forma e il colore degli organi della pianta sfruttati dall'uomo (ad es. semi, frutti e foglie), e tratti legati alla dispersione del seme (ad es. la frantumazione e la dormienza). Nonostante la dimensione del seme e del frutto siano le differenze più evidenti tra la forma domesticata e quella selvatica, la perdita di meccanismi di dispersione del seme rappresenta infatti uno dei fattori maggiori che hanno ridotto la fitness delle piante domesticate nell'ambiente selvatico, evitando che queste si riproducessero fuori dall'ecosistema agricolo.

Analizzando la domesticazione dalla prospettiva della genetica di popolazioni, un evento di domesticazione porta ad una riduzione della diversità genetica e ad un aumento della divergenza tra la forma selvatica e la forma domesticata, sia a causa di fattori demografici che influenzano l'intero genoma, sia a causa di selezione a livello di loci target (Glèmin and Bataillon 2009). Le specie allogame come il mais (*Zea mays*) sono caratterizzate generalmente da un effetto "collo di bottiglia" meno intenso rispetto alle specie autogame come il fagiolo (*Phaseolus vulgaris*) (Bitocchi et al. 2013). In particolare, studi precedenti hanno mostrato come in specie autogame, come la soia (*Glycine max*) e il riso (*Oryza sativa* ssp *japonica*) (Lam et al. 2010; Xu et al. 2012), si siano verificate riduzioni nella diversità genetica come conseguenza della domesticazione.

Segnali di selezione durante la domesticazione sono stati identificati nel 2-4% dei geni espressi in mais (S. I. Wright et al. 2005) e nel 7,6% del genoma di mais (Hufford et al. 2012). Questi studi suggeriscono un ruolo importante dell'effetto combinato di selezione, deriva e riduzione della ricombinazione nei loci associati fisicamente ai geni bersaglio di selezione. Un forte effetto *hitchiking* (Maynard Smith and Haigh 1974) è stato inoltre suggerito per il riso (Lu et al. 2006) e per il fagiolo (Papa et al. 2007), supportando la visione che la domesticazione nel suo insieme abbia avuto un impatto sul genoma maggiore rispetto agli effetti che possono essere spiegati unicamente dall'azione della selezione.

Le tecnologie di sequenziamento di nuova generazione offrono un'occasione unica per l'analisi del genoma: non solo per ottenere informazioni sul genotipo, ma anche per analizzare il fenotipo molecolare del genoma attraverso l'analisi del trascrittoma, del metaboloma e del proteoma. Studi recenti hanno osservato rilevanti cambiamenti nell'espressione del trascrittoma nel mais, senza riduzione nella diversità di espressione dei geni (Hufford et al. 2012; Swanson-Wagner et al. 2012). Un'estensione di questi studi ad altre colture permetterebbe di indagare meglio le conseguenze della domesticazione sull'intero genoma.

In questo studio è stata studiata la domesticazione del fagiolo comune nel Mesoamerica, con gli scopi principali di 1) descrivere i cambiamenti genomici legati alla domesticazione usando RNA-seq e 2) identificare le varianti molecolari nel genoma del fagiolo comune responsabili delle variazioni fenotipiche alla base del processo di domesticazione. L'analisi dei network di espressione non verrà trattata in questo capitolo, ma è disponibile nell'articolo pubblicato (Bellucci et al. 2014).

Nella storia evolutiva di *Phaseolus vulgaris* (2n=2x=22) si sono verificati almeno due eventi di domesticazione, uno nel Mesoamerica e uno nelle Ande (come osservato in (Bitocchi et al. 2013)). Questi eventi paralleli di domesticazione e la domesticazione di 4 specie di *Phaseolus* evolutivamente vicine rendono il fagiolo comune un modello unico e importante per studiare la domesticazione e l'evoluzione della coltivazione.

Materiali e metodi

Preparazione del dataset

Il mio contributo a questo studio è legato principalmente all'analisi dei dati, ma per completezza descriverò brevemente anche la preparazione del dataset, svolta presso il gruppo di Roberto Papa del Dipartimento di agricoltura, cibo e scienze ambientali dell'Università Politecnica delle Marche e dal gruppo di Massimo Delledonne del Dipartimento di biotecnologia dell'Università di Verona.

21 genotipi inbred sono stati selezionati a partire da una collezione rappresentativa di genotipi di *Phaseolus vulgaris* (Rossi et al. 2009; Nanni et al. 2011; Bitocchi et al. 2012; Desiderio et al. 2012), al fine di massimizzare la diversità genetica nel campione. Il campione raccolto è costituito da 10 genotipi mesoamericani selvatici (MW), 8 mesoamericani domestici (MD), 2 domesticati andini (AD) e 1 selvatico andino (AW). Quattro di questi campioni (3 mesoamericani, 1 andino) catturavano la maggior parte della diversità genetica, e sono stati in seguito utilizzati per la costruzione di un trascrittoma di riferimento (Tabella 1, tratta dal lavoro pubblicato). Questo approccio è stato preferito rispetto ad uno basato su un genoma di riferimento in quanto il genoma di riferimento esistente all'epoca dell'inizio di questo lavoro era basato su un genotipo andino, e avrebbe potuto portare a una perdita di regioni informative a causa della divergenza nota tra le due popolazioni.

	Accession number/name	Population code[1]	Source[2]	Country
1	G12873*	MW	CIAT	Mexico
2	G9989	MW	CIAT	Mexico
3	G11050	MW	CIAT	Mexico
4	G12979*	MW	CIAT	Mexico
5	G22837	MW	CIAT	Mexico
6	G24378*	MW	CIAT	Mexico
7	PI325677	MW	USDA	Mexico
8	PI417770	MW	USDA	Mexico
9	G20515	MW	CIAT	Mexico
10	G12922	MW	CIAT	Mexico
11	G5191	MD	CIAT	Venezuela
12	PI151017	MD	USDA	Chile
13	PI201349	MD	USDA	Mexico
14	PI281981	MD	USDA	Mexico
15	PI300668	MD	USDA	Chile
16	PI309831	MD	USDA	Costa Rica
17	PI310660	MD	USDA	Guatemala
18	PI311794	MD	USDA	El Salvador
19	W617475*	AW	USDA	Argentina
20	Midas	AD	P.Gepts	Argentina
21	PI298109	AD	USDA	Brazil

Tabella 1. Genotipi utilizzati in questo studio. [1]MW, Mesoamericano selvatico; MD, Mesoamericano domesticato; AW, Andino selvatico; AD, Andino domesticato. 2 CIAT, Centro Internazionale di Agricoltura Tropicale; USDA, Dipartimento di Agricoltura degli Stati Uniti; il genotipo Midas è stato gentilmente fornito dal Prof. Paul Gepts di UC Davis, USA. * Genotipi usati per l'assembly de novo del trascrittoma di riferimento.

I 21 genotipi sono stati cresciuti in serra in condizioni controllate (umidità relativa ca. 70%, temperatura media 25°C), ed è stata raccolta per ognuno di essi la prima foglia trifogliata espansa nella fase stazionaria per minimizzare differenze nei livelli di espressione genica dovute a un diverso stadio di sviluppo degli individui. La foglia è stata congelata e conservata in azoto liquido.

3 µg di RNA per ogni campione sono stati utilizzati per preparare una libreria Illumina RNA-seq non direzionale, usando kit di preparazione TruSeq RNA v2 (Illumina). La libreria è stata quantificata con PCR quantitativa, e analizzata con un chip bioanalyzer DNA 1000 serie II (Agilent) per valutarne la qualità. Il sequenziamento è stato eseguito con Illumina HiSeq 1000 Sequencer utilizzando i

kit TruSeq SBS v3-HS e TruSeq PE cluster v3-cBot-HS (Illumina), generando una libreria di *reads* paired-end di 100 bp.

Le *reads* dei quattro campioni rappresentativi della variabilità genetica sono state assemblate *de novo* utilizzando Trinity v R0211-11-2 (Grabherr et al. 2011) su ogni singolo campione. I contig ottenuti sono stati filtrati in modo da mantenere l'isoforma più lunga di ogni gene. I contig filtrati di ogni campione sono stati uniti con i contig degli altri campioni, e sono stati rimossi i contig ridondanti (identità maggiore del 90%) utilizzando CD-HIT_EST (Weizhong Li and Godzik 2006). I contig ottenuti sono stati confrontati con le sequenze del database TAIR 10 di *Arabidopsis thaliana* utilizzando BLASTX (Altschul et al. 1997) con un E-value < 10E-2.

Le *reads* di ognuno dei 21 genotipi sono state mappate sul trascrittoma di riferimento utilizzando bwa v 0.6.2 r126 (H. Li and Durbin 2009) con parametri di default e una qualità di mapping minima di 30 per minimizzare errori di allineamento e *reads* mappate in posizioni multiple. I siti variabili sono stati chiamati utilizzando Samtools 0.1.18 (H. Li et al. 2009) e VarScan v 2.2.8 (Koboldt et al. 2012) con un p-value massimo di 0,01 e un coverage minimo di 3 *reads* per non penalizzare trascritti presenti in piccole quantità nei campioni. Solamente posizioni nel trascrittoma coperte da almeno 3 *reads* in tutti i 18 genotipi mesoamericani analizzati sono state considerate per la chiamata delle varianti. In tutte le posizioni in cui uno SNP omozigote (percentuale di *reads* che supportano l'allele alternativo superiori al 75%) è stato chiamato in almeno un campione, sono stati analizzati i campioni in cui in quella posizione non sono stati chiamati SNP, adottando i seguenti criteri: 1) se il numero di *reads* che coprono il sito è maggiore o uguale a 3 e la percentuale di *reads* che supportano la base della referenza è maggiore del 75%, è stata chiamato un sito omozigote per l'allele del trascrittoma di riferimento; 2) se il numero di *reads* che coprono il sito è maggiore o uguale a 3 e la percentuale di *reads* che supportano la base dell'allele alternativo (allele già chiamato in altri campioni con un P-value<0.01) è maggiore del 75%, è stato chiamato un sito omozigote per la base alternativa; 3) se il numero di *reads* che coprono il sito è maggiore o uguale a 3 e la percentuale di *reads* che supportano la base dell'allele alternativo (allele già chiamato in altri campioni con un P-value<0.01) è compresa tra il 25% e il 75%, è stata chiamato un sito eterozigote.

Le posizioni in cui non è stato possibile determinare il genotipo in almeno 15 campioni mesoamericani sono state rimosse, permettendo un massimo di 3 individui mancanti per sito. Nelle analisi successive sono stati utilizzati solamente siti biallelici omozigoti.

Una piccola frazione, pari al 2,73% delle basi, risultava mancante nel dataset di 188.107 SNPs, in frazioni più o meno simili nei vari gruppi di individui. I dati mancanti sono stati quindi imputati utilizzando l'algoritmo di clustering implementato in *fastPhase1.4* (Scheet and Stephens 2006), un metodo che non richiede informazioni di pedigree e utilizza l'informazione della popolazione. Gli individui sono stati assegnati a tre gruppi: individui mesoamericani domesticati, individui mesoamericani domesticati e andini. Non è stata fatta una ricostruzione di aplotipi, e i genotipi sono stati imputati indipendentemente per ogni contig, impostando il numero di cluster a 15. La linea di comando utilizzata con *fastPhase* contiene i seguenti parametri: –KL1 –KU15 –Ki2 –H4 –n –B –u.

Per tutti i trascritti è stata eseguita un'analisi BLASTX (Altschul et al. 1997) contro il dataset proteico TAIR10 di *Arabidopsis thaliana*. Inoltre, per caratterizzare la funzione dei contigs neutrali o selezionati positivamente, è stata condotta una MapMan GSEA (Gene Set Enrichment Analysis).

I livelli di espressione dei geni sono stati valutati con TopHat2 (Kim et al. 2013) e HTSeq (Anders and Huber 2010). I contigs con forti differenze nei livelli di espressione tra i genotipi mesoamericani domesticati e i genotipi mesoamericani selvatici sono stati identificati con DESeq v 1.6.1 (Anders and Huber 2010) ($|\log(FC)|>1$; FDR<5%). Il coefficiente di variazione è stato calcolato come il rapporto tra la deviazione standard e il numero medio di frammenti che mappavano in ogni contig, per ogni genotipo, nei campioni selvatici e domesticati.

Analisi di genetica di popolazioni

Una prima analisi esplorativa delle relazioni genetiche degli individui è stata fatta basandosi su uno scaling multidimensionale metrico utilizzando la funzione *cmdscale* dell'ambiente statistico R (R Core Team 2015). Le distanze genetiche sono state calcolate come 1 meno la frazione media di alleli condivisi.

Il numero totale e medio di siti segreganti (S), l'eterozigosi attesa (He, (Nei 1978)), il numero di aplotipi (nH), il numero medio di differenze a coppie (π, (Tajima

1983)) e lo stimatore di theta di Watterson (θ, (G A Watterson 1975)) sono state calcolate sugli SNPs di ogni contigs utilizzando *Arlequin 3.5* (Laurent Excoffier and Lischer 2010). La divergenza tra la popolazione mesoamericana domesticata e quella selvatica è stata valutata utilizzando ϕ_{ST} (L. Excoffier, Smouse, and Quattro 1992).

Per valutare la riduzione di diversità nella popolazione mesoamericana domesticata rispetto alle selvatica è stata utilizzata la statistica di riduzione di diversità proposta da (Vigouroux et al. 2002) nella formulazione $1-(X_{MD}/X_{MW})$, dove X_{MD} è la diversità nella popolazione domesticata mesoamericana e X_{MW} la diversità nella popolazione selvatica mesoamericana. La diversità è stata calcolata con tre diverse statistiche: He, π e θ. Questa statistica è compresa tra 0 e 1, dove 0 indica nessuna riduzione di diversità e 1 una riduzione totale di diversità. La differenza tra le distribuzioni di ognuna delle statistiche di diversità genetica (S, He, nH, π e θ) nelle popolazioni selvatica e domesticata è stata valutata con il test dei ranghi di Wilcoxon per dati appaiati.

Identificazione dei contigs putativamente sotto selezione

I contigs selezionati positivamente nella popolazione mesoamericana domesticata rispetto alla popolazione selvatica sono stati identificati calcolando due indici di selezione e testando la loro significatività usando delle simulazioni. Gli indici di selezione sono basati su due statistiche di variazione genetica tra gruppo ed una entro gruppo, e la loro distribuzione neutrale (in assenza di selezione) è stata ottenuta con simulazioni in coalescente utilizzando le informazioni demografiche provenienti da studi precedenti in *P. vulgaris*.

La distribuzione neutrale delle statistiche riassuntive è stata ottenuta con simulazioni in coalescente, utilizzando tre diversi scenari di domesticazione verosimili ricostruiti sulla base di studi precedenti ((Mamidi et al. 2011; Mamidi et al. 2013), Figura 1).

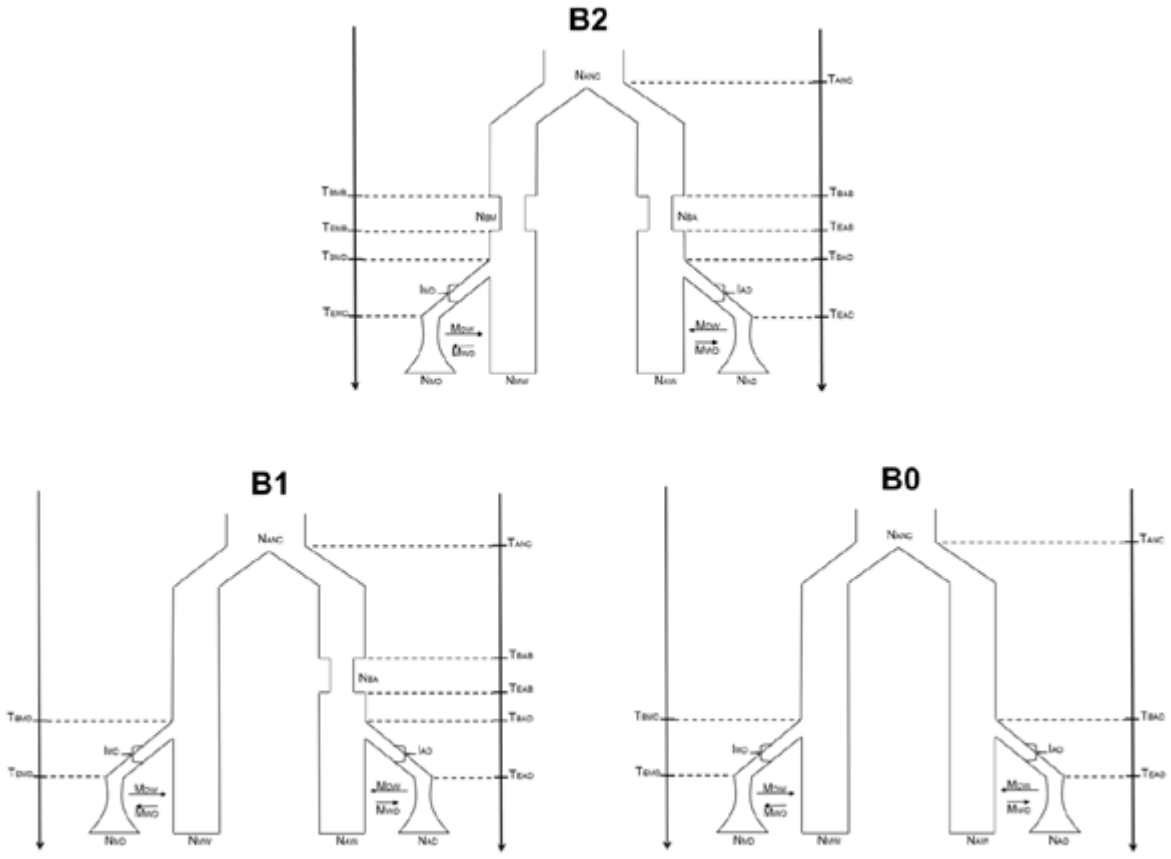

Figura 1. Modelli demografici utilizzati in questo studio.

In tutti i modelli, il pool genico andino e mesoamericano si originano da una popolazione ancestrale, e la domesticazione è un evento indipendente in ognuno dei due pool. Solo in tempi più recenti i gruppi domesticati si espandono esponenzialmente, ed alcuni eventi di ibridizzazione avvengono tra individui domesticati e selvatici. I modelli B1 e B2 includono un collo di bottiglia nel gruppo andino o in entrambi gruppi, rispettivamente.

Per ogni parametro abbiamo definito delle distribuzioni a priori (Tabella S1) basate sul tasso di incertezza di stime precedenti (Mamidi et al. 2011; Mamidi et al. 2013). Per ogni modello sono state eseguite 100000 simulazioni campionando a caso combinazioni di valori dalle distribuzioni a priori utilizzando il programma ABCsampler dal pacchetto ABCtoolbox (Wegmann et al. 2010). In ogni simulazione abbiamo generato un campione per ogni popolazione della stessa dimensione di quello osservato (10 individui aploidi per la popolazione

mesoamericana selvatica, 8 per quella mesoamericana domesticata, 2 per la popolazione andina domesticata e uno per la popolazione andina selvatica). Le lunghezze delle sequenze di DNA simulate sono state estratte da una distribuzione calibrata sulla base della distribuzione della lunghezza dei contigs nei campioni reali.

Abbiamo calcolato tre statistiche potenzialmente affette dalla selezione diversificante nella popolazione domesticata rispetto alla selvatica per ogni locus. La prima, la statistica di differenziazione di popolazioni (ϕ_{ST}, (L. Excoffier, Smouse, and Quattro 1992)), è stata calcolata tra le due popolazioni, seguendo la visione classica che loci sottoposti a diverse pressioni selettive in diverse popolazioni possono essere individuati come outlier quando si confrontano le popolazioni (Lewontin and Krakauer 1973). La seconda, la statistica "*locus specific branch length*" (LSBL, (Shriver et al. 2004)), è basata sulla distanza genetica tra le due popolazioni tenendo in considerazione anche un terzo gruppo (il gruppo andino), per identificare quale delle due popolazioni è stata sottoposta a pressione selettiva positiva. La terza è stata ideata per catturare le differenze relative nei livelli di variabilità dovuti alla selezione (Pritchard, Pickrell, and Coop 2010) andando a misurare il valore assoluto della differenza tra le variabilità genetiche nei gruppi domesticato e selvatico e utilizzando la somma della variabilità per standardizzare. Questa statistica è stata calcolata come il rapporto tra $|S_{MW}-S_{MD}|$ e $S_{MW}+S_{MD}$, dove S_{MW} è il numero di siti segreganti della popolazione selvatica e S_{MD} quello della popolazione domesticata. Tutte queste tre statistiche aumentano all'aumentare dell'evidenza di selezione.

ϕ_{ST} e il numero di siti segreganti sono stati calcolati utilizzando la versione a linea di comando di Arlequin 3.5 (Laurent Excoffier and Lischer 2010), e normalizzati utilizzando le distribuzione neutrali ottenute tramite simulazione.

Le tre statistiche sono state combinate in due diversi indici. Il primo è composto dalla somma di tutte le statistiche ed è stato calcolato per 26.116 contigs, in quando la statistica basata sul numero di siti segreganti non può essere calcolata nei 1.127 contigs in cui alleli diversi sono fissati nei due gruppi. Per questi contigs abbiamo creato un secondo indice sommando solamente ϕ_{ST} e la statistica LSBL standardizzati. La stessa procedura è stata utilizzata nei dati simulati per generare delle distribuzioni di questi indici assumendo che il pattern di variabilità sia stato

determinato solamente da fattori demografici, non selettivi. Il p-value per ogni contig è stato poi calcolato come la frazione di valori dell'indice simulati più grandi del valore reale, e corretto per falsi positivi usando l'approccio di Benjamini-Hochberg (Benjamini and Hochberg 1995) implementato nella funzione *p.adjust* nell'ambiente statistico R (R Core Team 2015). Questo approccio è stato ripetuto per ognuno dei tre modelli, ottenendo tre diverse liste di p-value corretti. Abbiamo definito come contigs selezionati positivamente quei contigs dove il p-value corretto era inferiore a 0,05 per ognuno dei tre modelli.

Risultati

Sequenziamento del trascrittoma e assemblaggio

Il sequenziamento del trascrittoma ha generato in media 38 milioni di *reads* paired-end (100bp x 2) per campione. Utilizzando Trinity per l'assemblaggio *de novo* dei quattro campioni rappresentativi sono stati ottenuti da 55.069 a 70.826 cluster di contigs, come definito dal modulo Chrysalis di Trinity, dove ogni cluster rappresenta un singolo gene. Il contig più lungo di ogni cluster è stato scelto come rappresentativo, e le ridondanze tra i quattro genotipi sono state eliminate con CD-HIT-EST ottenendo un set di 124.166 sequenze (geni condivisi e geni unici dei quattro genotipi). Queste sequenze sono state usate come trascrittoma di riferimento per le analisi successive.

Il mapping delle sequenze di ogni campione sul trascrittoma di riferimento ha portato alla chiamata di 284.812 polimorfismi a singolo nucleotide (SNPs) di alta qualità su 43.789 contigs. Contigs con SNPs eterozigoti o con inserzioni/delezioni non sono stati considerati. Da questo dataset sono state eliminate le posizioni mancanti in più di tre genotipi Mesoamericani e le posizioni con più di due alleli, ottenendo così un dataset finale di 188.107 SNPs in 27.243 contigs. Considerando solamente gli individui del Mesoamerica i contigs polimorfici sono 26.141, con 25 di questi fissati per stati allelici alternativi tra le popolazioni domesticate e selvatiche (Tabella 2).

SNPs biallelici	188.107
Contigs totali	27.243
Contigs monomorfici nel campione mesoamericano	1.102
Contigs polimorfici nel campione mesoamericano	26.141
Contigs monomorfici in entrambe le popolazioni (selvatica e domesticata) mesoamericane, ma per alleli alternativi	25
Contigs polimorfici condivisi tra la popolazione selvatica e domesticata	13.411
Contigs monomorfici nella popolazione domesticata, polimorfici nella popolazione selvatica	12.014
Contigs monomorfici nella popolazione selvatica, polimorfici nella popolazione domesticata	691

Tabella 2. Numero di SNPs e contigs individuati in questo studio.

Analisi di struttura, diversità ed espressione

Le analisi di scaling multidimensionale (MDS, Figura 2A) riproducono le struttura genetica conosciuta delle popolazioni di fagiolo comune: i pool mesoamericani e andini sono separati, così come sono separate le forme domesticate e selvatiche. L'analisi rivela anche che, rispetto alla popolazione domestica, la popolazione selvatica è caratterizzata da un maggior livello di diversità (i punti sono più dispersi). Il risultato è concorde con le statistiche stimate (S, nH, π, θ e He) (Tabella 3), con il 60% di perdita di variabilità nella popolazione domesticata. Inoltre circa la metà dei contigs polimorfici nell'intero Mesoamerica (46%) risultano monomorfici nella popolazione domesticata (Tabella 2). La differenza di variazione genetica tra la popolazione domesticata e quella selvatica è altamente significativa per tutti gli indici (Wilcoxon test, p-value <=2,2 10E-16). (Figura 2B)

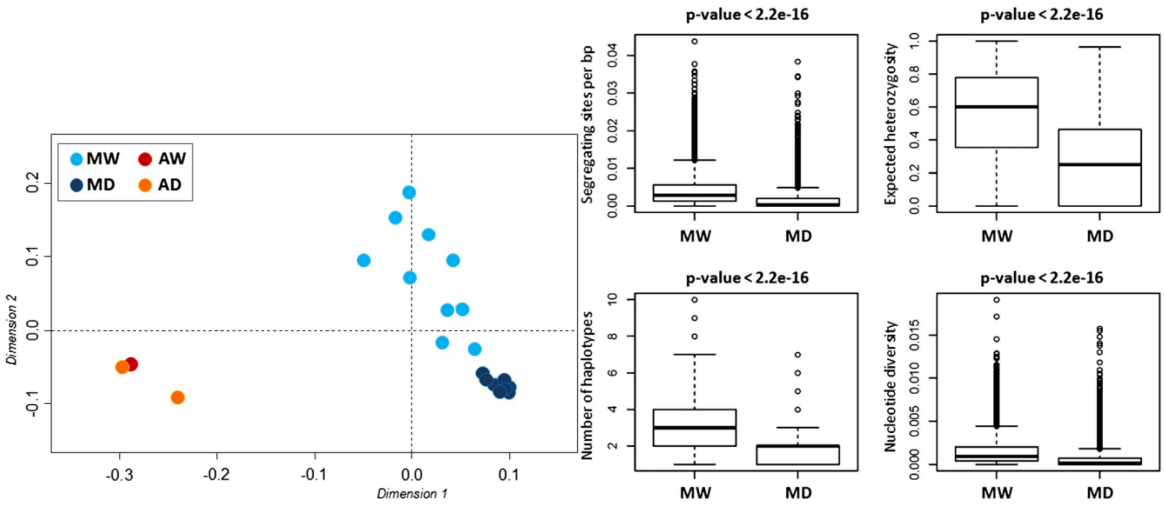

Figura 2. Rappresentazione delle relazioni genetiche tra I 21 genotipi di fagiolo comune ottenuta con un'analisi di scaling multidimensionale (A). Box plots del numero di siti segreganti (per coppia di basi), dell'eterozigosi attesa, del numero di aplotipi e della diversità nucleotidica nelle popolazioni mesoamericane domesticate (MD) e selvatiche (MW), valutate in tutti i contig. La significatività statistica è stata calcolata con il test dei ranghi di Wilcoxon per dati appaiati (p-value: sopra ogni box plot). (B)

Statistica	MW	MD
N	10	8
S	153.971	56.053
S_1	5,9	2,1
He	0,57	0,25
nH	3,5	1,8
π	2,11	0,85
θ	2,08	0,83
ΔH	0,56	
ΔH_1	0,56	
L_π	0,60	
$L_{\pi 1}$	0,58	
L_θ	0,60	
$L_{\theta 1}$	0,58	
ϕ_{ST}	0,15	

Tabella 3. Stime di diversità nelle popolazioni mesoamericane selvatica (MW) e domesticata (MD) calcolate considerando i contigs polimorfici nell'intero campione mesoamericano (264141 contigs) e stime di perdita di diversità nucleotidica. Legenda: N, dimensione campionaria; S, numero totale di siti segreganti; S_1, numero medio di siti segreganti per contig; He, eterozigosi attesa media; nH, numero medio di aplotipi per contig; π e θ, stime medie di diversità nucleotidica; ΔH, L_π e L_θ, perdita di diversità nucleotidica nel confronto MW ed MD calcolata da He, π e θ medie; ΔH_1, $L_{\pi 1}$ e $L_{\theta 1}$, perdita di diversità nucleotidica nel confronto MW ed MD calcolata da He, π e θ medie per contig; ϕ_{ST} medio.

L'analisi dell'espressione differenziale tra la popolazione mesoamericana selvatica e quella domesticata, considerando i diversi genotipi come repliche, ha permesso

di individuare 198 contigs su 27.243 (lo 0,7%) espressi in maniera differenziale nelle due popolazioni. 146 di questi (il 74%) hanno livelli di espressione inferiori nella popolazione domesticata. Inoltre, il cambiamento logaritmico nei livelli di trascrizione tende significativamente a valori negativi (media -0,09, mediana -0,02, *skewness* -3,49), segnale di un abbondanza di bassi livelli di trascrizione nella popolazione domesticata, con il cambiamento logaritmico medio significativamente più piccolo di 0 (p-value del test di Wilcoxon a due code: P<= 2E-16). Il coefficiente di variazione dell'espressione genica è più alto nella popolazione selvatica (0,57) che in quella domesticata (0,47), mostrando quindi una riduzione del 18% nella popolazione domesticata rispetto alla selvatica.

Selezione

2.364 contigs (9%) sono stati identificati come target di selezione diretta (o indiretta, dovuta ad *hitchiking*) durante la domesticazione utilizzando la simulazione delle dinamiche evolutive delle popolazioni mesoamericana e andina. La maggior parte dei contigs selezionati (82%) sono fissati nella popolazione domesticata e polimorfici nella popolazione selvatica, con il 14,2% dei contigs che mostrano polimorfismi condivisi tra le due popolazioni. Una piccola frazione (2,8%) è rappresentata da contigs fissati nella popolazione selvatica e polimorfici in quella domesticata. L'1% rimanente risulta fissato in entrambe le popolazioni per stati allelici alternativi.

I contigs differenzialmente espressi nella popolazione selvatica rispetto alla domesticata sono significativamente arricchiti di contig selezionati positivamente rispetto ai contig neutrali (2,75% vs 0,53%, con un arricchimento di circa 4 volte). Questo dato suggerisce come la selezione sia attiva già a livello del pattern di espressione della prima vera foglia. In parallelo, la perdita di diversità di espressione dovuta alla domesticazione sembra essere significativamente più alta (P<0,0001, test del chi-quadro) per i contigs selezionati positivamente (26%) rispetto ai contigs neutrali (17%). Questo effetto potrebbe essere il risultato della selezione diretta o di effetto *hitchiking* nelle regioni regolative (all'interno o all'esterno dell'esoma).

Le analisi di arricchimento di set genici (GSEA) mostrano come singoli MapMan bins (categorie che descrivono processi nella pianta) arricchiti non siano comuni

per i set neutrali e selezionati, indicativo del fatto che i due set abbiano funzioni differenti o partecipino a vie metaboliche diverse. I contig selezionati positivamente sono arricchiti per geni che in *Arabidopsis thaliana* sono implicati nella regolazione della trascrizione dell'RNA, nella sintesi di proteine ribosomali, nel processa mento/regolazione dell'RNA e nella riparazione di DNA. Al contrario, i contig neutrali hanno un numero più basso di MapMan bins arricchiti, bins che includono proteine implicate nei cofattori e nel metabolismo delle vitamine e implicate nel metabolismo dei nucleotidi.

In aggiunta, l'informazione funzionale permette di capire se un sottoinsieme dei contigs identificati come selezionati sia associato al processo di domesticazione in altre specie. I 380 contigs con l'indice di selezione più alto, includendo anche i 23 contigs con uno stato allelico alternativo fissato tra la popolazione domesticata e selvatica e i 67 contigs monomorfici nella popolazione selvatica e polimorfici in quella domesticata sono stati analizzati singolarmente. Questa analisi mostra come molti contigs selezionati positivamente siano omologhi a geni implicati nel processo di domesticazione in altre specie o abbiano funzioni associate con la domesticazione, come la risposta alla luce, la trasduzione del segnale, lo sviluppo della pianta e lo stress biotico e abiotico. Per esempio, tra i geni annotati selezionati positivamente che mostrano un livello maggiore di diversità genetica nella popolazione selvatica rispetto alla domesticata vi è una sequenza omologa a GIGANTEA (GI), un gene implicato nella risposta al fotoperiodo, che regola la fioritura in maniera controllata da un orologio circadiano. In *Arabidopsis*, nelle giornate lunghe, GI agisce a monte di un pathway aumentando l'abbondanza di mRNA di CONSTANSE (CO) e FLOWERING TIME (FT). CO e FT sono due target di selezione noti in riso (Takahashi and Shimamoto 2011; Wu et al. 2013) e girasole (*Helianthus anuus*, (Blackman et al. 2011)). Nel pisello (*Pisum sativum*) (Hecht et al. 2007) LATE BLOOMER1 (LATE1) è stato identificato come l'ortologo in pisello di GI in *Arabidopsis* ed è necessario per la promozione della fioritura, la produzione di uno stimolo della fioritura mobile e l'induzione di un omologo di FT in condizione di lunghe giornate. Un altro esempio interessante tra i contigs selezionati positivamente con due stati allelici alternativi è l'omologo di YABBY5 (YAB5), un fattore di trascrizione implicato nella regolazione della frammentazione dei semi nelle specie di cereali, inclusi il sorgo (*Sorghum bicolor*), riso e mais (Lin et al. 2012). Un fattore di trascrizione YAB-lke (FASCIATED) ha effetto sul numero

di carpelli durante la fioritura e/o lo sviluppo del frutto nel pomodoro (*Solanum lycoperiscum*, (Cong, Barrero, and Tanksley 2008)).

Tra i 67 contigs che mostrano un aumento di variabilità nella popolazione domesticata c'è un omologo di K+ UPTAKE TRANSPORTER6 (KUP6). (Osakabe et al. 2013) hanno dimostrato che la famiglia di trasportatori di potassio KUP ha un ruolo importante nella risposta allo stress da carenza di acqua e nella crescita; inoltre i trasportatori di potassio di tipo KUP sono indotti da diversi stress che hanno una componente osmotica, e inibiscono specificamente l'espansione cellulare, aumentando la tolleranza alla siccità.

Discussione e conclusioni

Questo capitolo mostra gli effetti profondi che la domesticazione ha avuto sulla variazione genomica e sui pattern di espressione genica nel fagiolo comune. Circa 1 contig su 10 è stato verosimilmente selezionato durante la domesticazione: selezione direzionale in primo luogo, ma probabilmente anche selezione diversificante, con contigs del pool genico domesticato che presentano livelli di espressione diversi rispetto alla controparte selvatica. Le implicazioni pratiche per lo sviluppo futuro delle colture sono legate all'elevata variabilità nelle sequenze di DNA disponibili nel fagiolo selvatico, ma per poter sfruttare questa diversità della popolazione selvatica è necessario uno sforzo sostanziale per capire le complesse relazioni tra la diversità genotipica e fenotipica nelle popolazioni di piante. Come sottolineato in questo studio, nel fagiolo comune domesticato le regioni genomiche espresse hanno perso metà della diversità nucleotidica della popolazione selvatica durante la domesticazione in Mesoamerica. Rispetto al fagiolo comune, nel mais si è assistito ad una riduzione inferiore nei livelli di diversità nucleotidica (Hufford et al. 2012), suggerendo come l'effetto della domesticazione sulla diversità genetica del mais sia stato inferiore. Questi risultati contrastanti potrebbero essere spiegati con il diverso tipo di sistema di accoppiamento di queste due colture. Infatti nelle specie autogame, come il fagiolo, l'auto fertilizzazione comporta una riduzione della dimensione effettiva della popolazione, aumentando l'effetto della deriva genetica e l'estensione dei tratti in linkage disequilibrium. La conseguenza di questo fenomeno è la presenza di finestre genomiche più estese affette da sweep genetico (Glèmin and Bataillon 2009; Bitocchi et al. 2013), risultato

confermato dal ri-sequenziamento del riso (*O. sativa japonica*) e soia (Lam et al. 2010; Xu et al. 2012).

Questo studio ha inoltre dimostrato come ci sia stata una drastica riduzione, causata dalla domesticazione, nei pattern dell'espressione genica nell'intero set di geni. Un risultato simile è stato visto nel mais (Swanson-Wagner et al. 2012), anche se con intensità ridotta. Inoltre nel fagiolo comune si è osservato come la riduzione nella diversità sia presente anche in regioni implicate nella regolazione della trascrizione, dove circa il 20% di riduzione nel livello di espressione genica è stata associata alla domesticazione. In altre parole, si è dimostrato che la perdita di variabilità genetica ha delle estese conseguenze fenotipiche sulla diversità del trascrittoma. Nel caso del mais e del suo progenitore teosinte, nessuna riduzione nella variazione dell'espressione genica è stata osservata (Swanson-Wagner et al. 2012). E' rilevante notare come questi diversi pattern e livelli di espressione siano stati osservati in uno stadio di sviluppo che è considerato relativamente importante per la domesticazione, nonostante la presenza di foglie più larghe e sementi sia una caratteristica chiave della domesticazione (P. Gepts 2002).

La presenza nel fagiolo domesticato di trascritti per la maggior parte sotto espressi tra quelli differenzialmente espressi (74%) suggerisce che le mutazioni *loss-of-function*, meno frequenti rispetto a cambiamenti *gain-of-function*, siano una fonte di variazione disponibile che supporta la selezione durante rapidi cambiamenti ambientali (Olson 1999), come nel caso della transizione dall'ecosistema selvatico a quello coltivato. In supporto a questo, come osservato da Darwin (Darwin 1859), nelle piante domesticate i tratti legati alla domesticazione sono tratti recessivi (Lester 1989). Inoltre, è stato osservato un livello di espressione genica più basso nei trascritti domesticati rispetto a quelli selvatici, come se mutazioni leggermente deleterie (come perdita di funzione o riduzione dell'espressione) si siano accumulate nel pool genico domesticato a causa di *hitchiking*. Questo può essere considerato come il "costo della domesticazione". L'accumulo di mutazioni *loss-of-function* può anche essere stato causato da una riduzione nella ricombinazione, che potrebbe aver aumentato la frequenza di mutazioni deleterie nel pool domesticato influendo negativamente sulla fitness, come visto nel riso (Lu et al. 2006).

Circa il 10% dei contigs sono stati selezionati positivamente durante la domesticazione o sono in linkage fisico con geni selezionati. Questo supporta l'idea che la domesticazione abbia avuto un effetto rilevante nel genoma di fagiolo comune. Nel mais, allogamo, circa il 2-4% dei geni e il 7,6% del genoma è stato identificato come sotto selezione durante la domesticazione (S. I. Wright et al. 2005; Hufford et al. 2012). Nel girasole, prevalentemente allogamo, circa il 7,3% dei geni mostrano segni di selezione legati alla domesticazione (Chapman et al. 2008). Queste differenze potrebbero essere spiegate da un ruolo più rilevante dell'*hitchhiking* genetico nel fagiolo comune, legato al suo sistema di accoppiamento autogamo.

La maggior parte dei contigs selezionati durante la domesticazione mostra una riduzione nella diversità della popolazione domesticata rispetto a quella selvatica, come ci si attende dopo un evento di selezione positiva causato dalla domesticazione. E' stato anche osservato l'opposto, con il 2,8% dei contigs selezionati monomorfici nella popolazione selvatica e polimorfici in quella domesticata. Questo può essere spiegato dall'effetto della selezione diversificante, che ha aumentato i livelli di diversità funzionale nella popolazione domesticata. Le analisi funzionali del gene KUP6, legato alla resistenza alla siccità, mostrano come questo gene sia significativamente sovra espresso nella popolazione domesticata rispetto alla selvatica, come se la domesticazione avesse contemporaneamente aumentato sia la diversità funzionale dei geni selezionati, che la loro diversità genetica. Il dato suggerisce quindi che, in parallelo con una complessiva riduzione di diversità, la domesticazione abbia aumentato la diversità funzionale a specifici loci. Questo può essere imputato a mutazioni nuove (o a bassa frequenza) che sono state selezionate a causa dell'espansione della coltura in terreni con livelli non attesi di stress biotico o abiotico, o a causa di selezione di tratti che hanno migliorato l'uso degli organi della pianta da parte dell'uomo (De Alencar Figueiredo et al. 2008). I dati contribuiscono a risolvere il paradosso di Darwin (Dyer 1877; Glèmin and Bataillon 2009): la domesticazione è associata contemporaneamente a un aumento della diversità fenotipica per alcuni tratti, e ad una generale riduzione della variazione nucleotidica nei tratti sottoposti a selezione.

Questo lavoro presenta importanti implicazioni per lo sviluppo di strategie pre-accoppiamento. Analogamente ad altri studi, i risultati supportano la necessità di utilizzare germoplasmi selvatici per ulteriori miglioramenti della coltivazione e suggeriscono uno sforzo per la conservazione delle popolazioni selvatiche. Il lavoro mostra anche come l'effetto della domesticazione sia diffuso in tutto il genoma in termini di pattern di espressione e diversità, probabilmente a causa della combinazione di linkage e pleiotropia. Interazioni complesse tra geni e i loro livelli di espressione hanno giocato un ruolo chiave durante la domesticazione di questa specie, suggerendo come ulteriori miglioramenti genetici richiedano il contributo di nuovi strumenti provenienti dalla genomica, dalla fenotipizzazione molecolare e dalla fenomica. Infine, i risultati suggeriscono di considerare maggiormente la diversità nel pool domesticato (cioè le coltivazioni tradizionali), utilizzandola come fonte di nuova variazione genetica per il miglioramento delle colture.

Materiali supplementari

Tabella S1. Parametri demografici dei modelli B0, B1 e B2

Modello	Parametro	Descrizione	Unità	Distribuzione	Media	Dev. Std.	Min	Max
B0,B1,B2	TANC	Tempo di divergenza tra il pool genico Andino e Mesoamericano	Generazioni	Normale	111000	40000	67330	192835
B2	TBMB	Tempo dell'inizio del collo di bottiglia Mesoamericano	Generazioni	Normale	99833	750	98375	99833
B1,B2	TBAB	Tempo dell'inizio del collo di bottiglia Andino	Generazioni	Normale	98845	3000	94051	104027
B2	TEMB	Tempo della fine del collo di bottiglia Mesoamericano	Generazioni	Normale	67341	1400	64568	67341
B1,B2	TEAB	Tempo della fine del collo di bottiglia Andino	Generazioni	Normale	67858	1500	64655	70865
B0,B1,B2	TBMD	Tempo dell'inizio della domesticazione in Mesoamerica	Generazioni	Normale	8160	133	7922	8426
B0,B1,B2	TBAD	Tempo dell'inizio della domesticazione nelle Ande	Generazioni	Normale	8500	8	8495	8517
B0,B1,B2	TEMD	Tempo della fine della domesticazione in Mesoamerica	Generazioni	Normale	6260	150	5971	6567
B0,B1,B2	TEAD	Tempo della fine della domesticazione nelle Ande	Generazioni	Normale	7012	35	6945	7075
B0,B1,B2	NANC	Dimensione della popolazione effettiva ancestrale	Individui aploidi	Normale	418000	105000	266000	628000
B2	NBM	Dimensione della popolazione effettiva Mesoamericana durante il collo di bottiglia	Individui aploidi	Normale	168000	7500	153000	170000
B1,B2	NBA	Dimensione della popolazione effettiva Andina durante il collo di bottiglia	Individui aploidi	Normale	105000	20000	65000	142000
B0,B1,B2	NDM	Dimensione della popolazione effettiva domesticata Mesoamericana	Individui aploidi	Uniforme			100000	100000
B0,B1,B2	NWM	Dimensione della popolazione effettiva selvatica Mesoamericana	Individui aploidi	Normale	292000	240000	125000	773000
B0,B1,B2	NWA	Dimensione della popolazione effettiva selvatica Andina	Individui aploidi	Normale	137000	182000	70000	502000
B0,B1,B2	NDA	Dimensione della popolazione effettiva	Individui	Uniforme			100000	100000

		domesticata Andina	aploidi					
B0,B1,B2	IDM	Intensità del collo di bottiglia della domesticazione in Mesoamerica (percentuale)	Rapporto	Normale	47,65	3	41,66	52,13
B0,B1,B2	IDA	Intensità del collo di bottiglia della domesticazione nelle Ande (percentuale)	Rapporto	Normale	47,26	0,5	46,25	48,59
B0,B1,B2	MWD	Tasso di migrazione dalla popolazione selvatica alla domesticata	Tasso	Uniforme			0,000001	0,00001
B0,B1,B2	XM	Fattore di migrazione asimmetrico	Tasso	Uniforme			2	6
B0,B1,B2	MDW	Tasso di migrazione dalla popolazione domesticata alla popolazione selvatica	Tasso	XM*MWD				
B0,B1,B2	MU	Tasso di mutazione (per sito per generazione)	Tasso	Normale	0,000000001	0,000000005	1E-10	0,00000001
B0,B1,B2	L	Lunghezza del contig simulato	Paia basi	Normale	1300	1500	250	5000

Analisi di selezione in un dataset ristretto di geni di fagiolo comune

La metodologia per la ricerca di selezione è stata applicata ad un set ristretto di 43 geni noti per essere omologhi a geni di *Arabidopsis thaliana*, disponibili in letteratura. In questo nuovo studio, in fase di stesura, sono stati utilizzati due diversi modelli demografici basati su due diverse stime di parametri demografici, tra cui la nuova stima fatta grazie al nuovo genoma di riferimento di fagiolo (Schmutz et al. 2014), ed è stato modificato permettendo alla simulazione di gestire regioni introniche e frammenti di lunghezza fissa. Dopo aver corretto per test multipli, sono stati individuati 11 geni putativamente sotto selezione, alcuni dei quali già individuati in studi precedenti.

Questo studio è stato svolto in collaborazione con il gruppo di Roberto Papa dell'Università Politecnica delle Marche, che si è occupato della genotipizzazione dei geni e del calcolo delle statistiche di base.

Materiali e metodi

Sono stati utilizzati 43 genotipi di *P. vulgaris*: 19 da forme selvatiche del Mesoamerica (MW), 20 da forme domesticate del Mesoamerica (MD), 2 da forme domestiche delle Ande (AD) e 2 da forme selvatiche delle Ande (AW). Questi genotipi sono stati selezionate sulla base di una dettagliata caratterizzazione molecolare di un campione più ampio di *P. vulgaris*, rappresentativo dei differenti pool genici che caratterizzano la specie (Rossi et al. 2009; Nanni et al. 2011; Bitocchi et al. 2012; Bitocchi et al. 2013; Desiderio et al. 2012), allo scopo di massimizzare la diversità genetica del campione iniziale. Il DNA genomico è stato estratto per ogni genotipo da foglie giovani di una singola pianta cresciuta in serra usando il metodo di estrazione miniprep.

43 regioni geniche (da 150 a 900 bp) del genoma di fagiolo comune sono state utilizzate, 42 delle quali provenienti dalla letteratura (Hougaard et al. 2008; McConnell et al. 2010; Nanni et al. 2011; Goretti et al. 2014). Sulla base di analisi preliminari, che indicavano un comportamento da outlier del locus AN-Pv26, è stata sviluppata una nuova coppia di primers per amplificare una regione più ampia del gene corrispondente (gliceraldeide 3 fosfato deidrogenasi, GADPH), utilizzando la

regione genomica di questo gene in *Glycine max* per il disegno dei primers. E' stata ottenuta una regione a valle del gene originale (Goretti et al. 2014), sovrapposta con esso per 160 bp, e la nuova regione generata dall'unione delle due sequenze è stata utilizzata come singolo frammento.

Le sequenze di questi geni per i 43 genotipi di fagiolo comune utilizzati era già disponibile per i loci Leg044, Leg100, Leg133, Leg223 e PvSHP1 (Nanni et al. 2011; Bitocchi et al. 2012; Bitocchi et al. 2013). Le sequenze di 43 loci per 22 genotipi di *P.vulgaris* sono state ottenute dallo studio di (Goretti et al. 2014). Per i rimanenti 23 genotipi di *P.vulgaris* le sequenze sono state ottenute in questo studio, così come la sequenza del nuovo locus An-Pv 26. Alcune sequenze in alcuni individui risultavano di bassa qualità, e sono state quindi scartate. La lista completa dei loci è riportata in Tabella 2**.**

Analisi di diversità

L'allineamento e l'editing delle sequenze sono stati fatti utilizzando la versione 3.7 di MUSCLE (Edgar 2004) e la versione 7.0.9.0 di BIOEDIT (Hall 1999). Le inserzioni e le delezioni non sono state incluse nelle analisi. Le analisi di diversità sono state svolte con due raggruppamenti di campioni: la popolazione selvatica e la popolazione domesticata del Mesoamerica.

Sono stati calcolati i seguenti indici di diversità: il numero di siti variabili V, il numero di mutazioni η, il numero di siti informativi di parsimonia (Pi), il numero di siti polimorfici (S), il numero di aplotipi (H), la diversità aplotipica (Hd), π (Tajima 1983) e θ (G A Watterson 1975). Le stime sono state calcolate separatamente per le regioni codificanti e non codificanti (introni). La divergenza tra le popolazioni mesoamericane selvatica (MW) e domesticata (MD) è stata valutata calcolando il numero di mutazioni condivise ed uniche tra le due popolazioni all'interno di ogni pool genico e con la statistica F_{st} (Hudson, Slatkin, and Maddison 1992). Come proposto da Vigouroux (Vigouroux et al. 2002), per misurare la perdita di diversità nucleotidica tra le due popolazioni (MW e MD) è stata utilizzata la statistica $1-(X_{MD}/X_{MW})$, dove X_{MD} è la diversità nella popolazione domesticata mesoamericana e X_{MW} la diversità nella popolazione selvatica mesoamericana. La perdita di diversità nucleotidica è stata calcolata dalle stime medie di π e θ (L_{π} e L_{θ}, rispettivamente).

Il test non parametrico di Wilcoxon-Kruskal-Wallis è stato impiegato per testare le differenze tra le forme mesoamericane domesticate e selvatiche di *P.vulgaris* per le stime di diversità genetica π e θ, nelle regioni codificanti e nelle regioni non codificanti.

Analisi di selezione

Le regioni putativamente selezionate sono stati individuate seguendo lo stesso approccio usato nel capitolo precedente sull'intero trascrittoma (Bellucci et al. 2014). L'indice di selezione è stato calcolato separatamente nei frammenti esonici ed intronici (dove disponibili), e la sua significatività è stata testata utilizzando simulazioni in coalescente basate sulla storia demografica del fagiolo comune. I geni monomorfici (4 su 43) sono stati esclusi da questa analisi, mentre rispetto all'analisi precedente sono stati utilizzati tutti i campioni in esame, non solo i genotipi mesoamericani.

In sintesi, analogamente al precedente studio sul trascrittoma (Bellucci et al. 2014), l'indice di selezione è stato calcolato combinando tra statistiche normalizzate: l'F_{st} molecolare (L. Excoffier, Smouse, and Quattro 1992), la statistica "*locus specific branch length*" (LSBL,(Shriver et al. 2004)) e la differenza standardizzata tra la variazione genetica nella forma selvatica e domesticata. La prima e la seconda statistica permettono di catturare geni altamente differenziati tra il pool selvatico e domestico, mentre la terza è basata sul cambiamento atteso di diversità durante la selezione.

La distribuzione attesa in neutralità dell'indice di selezione è stata ottenuta con l'utilizzo di simulazioni grazie al pacchetto ABCtoolbox (Wegmann et al. 2010). Sono stati usati due diversi modelli, basati su precedenti ricostruzioni della storia demografica del fagiolo (modello 1, (Mamidi et al. 2011; Mamidi et al. 2013); modello 2, (Schmutz et al. 2014)) (Figura 1).

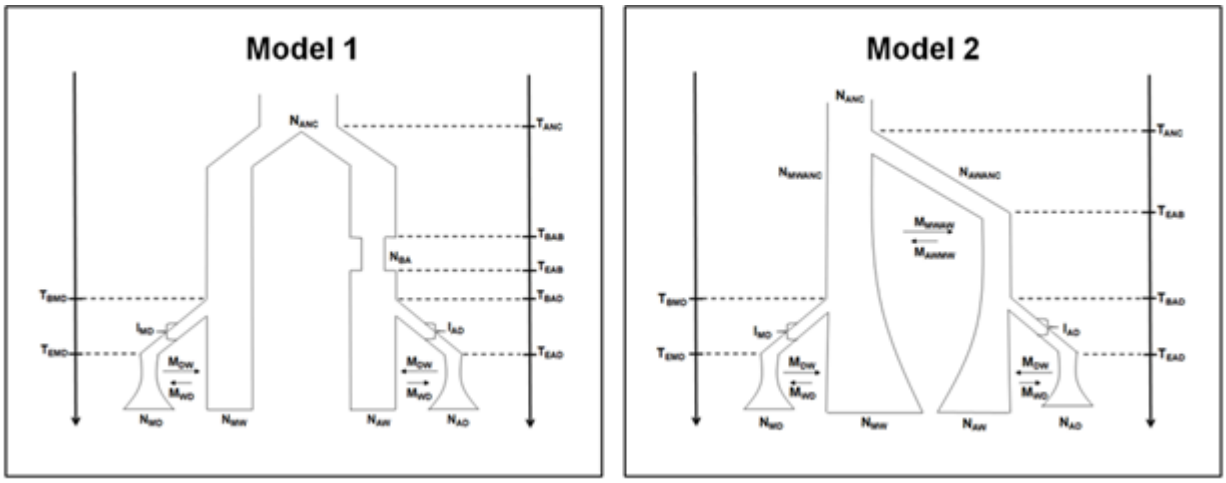

Figura 1. Modelli demografici utilizzati in questo studio, costruiti utilizzando informazioni da (Mamidi et al. 2011; Mamidi et al. 2013) e (Schmutz et al. 2014)

In entrambi i modelli le popolazioni mesoamericane e andine derivano da un pool ancestrale comune, con due eventi di domesticazione indipendenti dopo la divergenza in ognuna delle popolazioni. I due modelli differiscono nell'età dell'evento di collo di bottiglia associato con la colonizzazione delle Ande. Il modello 1 assume che queste evento sia recente, mentre il modello 2 assume la presenza di un collo di bottiglia subito dopo la separazione dal pool genico mesoamericano. Il modello 1 e il modello 2 assumo anche una diversa dinamica della crescita della popolazione in tempi recenti, con una crescita costante o esponenziale, rispettivamente. E' stata usata una distribuzione a priori di questi parametri per tenere in considerazione l'incertezza sulle stime (Tabella S1,Tabella S2).

Per ogni frammento esonico e intronico, le regioni con gap o dati mancanti sono state rimosse dall'allineamento, e sono stati analizzati unicamente i frammenti polimorfici. Campionando casualmente i parametri dei modelli dalle loro distribuzioni a priori sono state effettuate 100.000 simulazioni. La lunghezza dei frammenti simulati non è stata fissata con una distribuzione a priori, ma è stata specificata per ogni frammento sulla base della lunghezza del frammento nel dataset reale. La distribuzione a priori del tasso di mutazione utilizzata è diversa per le regioni esoniche ed introniche: un tesso di mutazione medio di 1,0 * 10E-08 per sito per generazione è stato utilizzato per gli introni, mentre per gli esoni è stato usato un

tasso 10 volte più basso (Bellucci et al. 2014). Il p-value per ogni frammento è stato calcolato come la frazione di indici simulati più grandi del valore osservato. I p-value sono stati corretti in seguito utilizzando l'approccio di Benjamini-Hochberg (Benjamini and Hochberg 1995) con la funzione *p.adjust* dell'ambiente R (R Core Team 2015), e la frazione di geni con un FDR (False Discovery Rate) più piccolo del 5% è stato considerata come putativamente sotto selezione.

I loci individuati come outliers sono stati valutati alla luce di lavori precedentemente pubblicati ((Bellucci et al. 2014), il tema del capitolo precedente, e (Schmutz et al. 2014)). La posizione dei 43 loci nel trascrittoma di riferimento di 27.243 contigs è stata valutata con Blast Like Alignment Tool (BLAT, (Kent 2002) v 34), implementato nell' iPlant Collaborative Web Portal (Oliver et al. 2013), con parametri di default. I risultati di BLAT sono stati filtrati per scegliere il miglior allineamento utilizzando il tool Best Hit for Blat Output v34, ottenendo un match per 25 loci con un E-value < 10E-24. Per confermare la procedura, l'analisi è stata ripetuta con BLASTN (Altschul et al. 1997) usando le impostazioni di default per allineare ogni locus con il rispettivo contig nella referenza. Questa procedura ha permesso un effettivo confronto dei loci analizzati con il trascrittoma di riferimento. In seguito i loci sono stati mappati nel genoma di riferimento (Schmutz et al. 2014) con lo stesso criterio, per permettere il confronto con le regioni genomiche e i geni identificate selezionati in entrambi i pool genici in (Schmutz et al. 2014).

Risultati

Un set di 43 frammenti genici è stato sequenziato in un campione di 43 genotipi di *P.vulgaris*, 39 dei quali forme selvatiche e domesticate del pool genico mesoamericano. Sono state ottenute sequenze di alta qualità per la maggior parte dei genotipi per tutti i loci, integrando i dati presenti in letteratura.

Quarantadue loci sono stati localizzati nel genoma di riferimento (Schmutz et al. 2014) attraverso un'analisi BLASTN; nessun match è stato trovato per il locus AN-Pv41. Con l'eccezione del cromosoma 3, i loci erano presenti in tutti e 10 i cromosomi rimanenti.

La porzione sequenziata per ogni locus ha una lunghezza compresa tra le 101 (AN-Pv35) e le 868 bp (Leg044), con una media di 439 bp per locus (Tabella 2). Il numero di basi totali analizzate per ogni genotipo è pari a 18.800 bp.

Analisi di diversità genetica

La diversità genetica è stata calcolata per ogni locus nella popolazione mesoamericana selvatica (MW) e nella popolazione mesoamericana domestica (MD). Il calcolo è stato ripetuto per le regioni codificanti e per gli introni. Un riassunto delle stime medie di diversità genetica calcolate è riportato nella Tabella 1.

Tutte le stime di diversità genetica sono più alte nella popolazione MW rispetto alla popolazione MD, sia nella porzione codificante, sia nella porzione intronica. In particolare, la stima di π è di 1,95 – 2,31 volte più alta in MW rispetto a MD rispettivamente negli esoni e negli introni. Lo stesso trend è stato osservato per θ, dove la stima è più alta nella popolazione MW di 1,74 (esoni) – 2,44 (introni) volte rispetto a MD. La differenza nelle stime di diversità genetica tra la popolazione selvatica e domesticata risulta significativa utilizzando il test non parametrico di Kruskal-Wallis ($P < 0.02$ per la porzione esonici, $P < 0,005$ per la porzione intronica).

Le stime di perdita di diversità, calcolate considerando le stime medie di diversità π e θ (Tabella 2), indicano chiaramente una forte riduzione di diversità della sequenza (circa il 50%) della forma domestica rispetto alla selvatica durante la domesticazione. Inoltre la perdita di diversità genetica è leggermente alta nelle regioni non codificanti (L_π=0,55 e L_θ=0,57) rispetto alle codificanti (L_π = 0,49 e L_θ=0,43).

Le stime di F_{st} e il pattern di mutazioni uniche e condivise in MW e MD medie sono riportate in Tabella 2. La stima di F_{st} media tra MW e MD è pari a 0,14 per la regione codificante e 0,17 non codificante. Il numero totale di mutazioni condivise tra MW e MD è pari a 70 nella regione codificante, 56 nella regione intronica. Il numero di mutazioni polimorfiche in MW e monomorfiche in MD sono 135 (64 negli esoni, 71 negli introni), mentre le mutazioni polimorfiche in MD e monomorfiche in MW sono solo 11 (8 nella regione codificante, 3 nella regione non codificante).

Porzioni codificanti (esoni)

Popolazione	N. loci	N	V	η	S	Pi	H	Hd	π x 10^{-3}	θ x 10^{-3}
MW	42	18,7	3,2	3,2	1,2	1,9	3,1	0,38	3,29	3,09
MD	42	19,8	1,9	1,9	0,7	1,2	2	0,2	1,69	1,77

Porzioni non codificanti (introni)

Popolazione	N. loci	N	V	η	S	Pi	H	Hd	π x 10^{-3}	θ x 10^{-3}
MW	32	18,7	6,8	7	1,9	4,9	4	0,46	7,63	7,61
MD	32	19,8	3,2	3,3	1,1	2,1	1,9	0,2	3,29	3,11

Tabella 1. Riassunto delle stime di diversità genetica calcolate per le popolazioni selvatica mesoamericana (MW) e domesticata mesoamericana (MD) nelle porzioni codificanti e non codificanti. N. loci, numero di loci utilizzati per il calcolo; N, dimensione campionaria media; V, siti variabili medi; η, numero totale di mutazioni; S, numero di siti segreganti; Pi, siti variabili informativi di parsimonia; H, numero di aplotipi; Hd, diversità aplotipica; π x 10^{-3} e θ x 10^{-3}, due misure di diversità nucleotidica da (Tajima 1983) e (G A Watterson 1975), rispettivamente.

	Differenziamento tra popolazioni (MW vs MD)				Perdita di diversità	
	F_{st}	SM	P_{MW}	P_{MD}	L_{π}	L_{θ}
Esoni	0,14	2,33	2,13	0,27	0,49	0,43
Introni	0,17	3,72	4,41	0,14	0,55	0,57

Tabella 2. Stime medie di F_{st} e pattern di mutazioni tra i genotipi selvatici e domesticati di fagiolo comune del Mesoamerica, considerando separatamente le regioni codificanti e non codificanti. SM, mutazioni condivise tra MW e MD; P_{MW}, mutazioni polimorfiche in MW ma monomorfiche in MD; P_{MD}, mutazioni polimorfiche in MW ma monomorfiche in MD; L_{π}, perdita di diversità nucleotidica calcolata su π (Tajima 1983); L_{θ} perdita di diversità nucleotidica calcolata su θ (G A Watterson 1975).

Analisi di selezione

I p-value associati ad ogni gene sono riportati in Tabella 2. Quattro loci (AN-Pv9, AN-Pv55, AN-Pv32 e ggsE20) risultano monomorfici nell'intero campione, e non sono stati utilizzati in questa analisi. Quando il Modello 1 è stato utilizzato per generare la distribuzione nulla dell'indice di selezione, 7 (17.9%) su 39 geni analizzati (AN-Pv22, AN-Pv26_1, AN-Pv33, AN-DNAJ, g523, Leg133, Leg223) hanno almeno un frammento esonico o intronico identificato come putativamente selezionato durante la domesticazione in Mesoamerica. Tre di loro mostrano un segnale di selezione unicamente nella regione intronica, tre unicamente nella regione esonica e uno in entrambi i tipi di regione. Quattro geni addizionali (AN-Pv64, AN-Pv69, AN-TGA e

PvSHP1) vengono identificati come putativamente selezionati quando viene utilizzato il Modello 2 (tre nella regione intronica, uno nella regione esonica); tutti i geni identificati con il Modello 1 sono inclusi nel set di loci identificati con il Modello 2.

					Modello 1		Modello 2	
Gene[1]	**Esoni[2]**	**Introni[2]**	**Lunghezza (esoni)**	**Lunghezza (introni)**	**FDR (esoni)**	**FDR (introni)**	**FDR (esoni)**	**FDR (introni)**
AN-Pv1	X	X	230	12	0,35	0,17	0,37	0,15
AN-Pv2	X	-	352	-	0,6	-	0,66	-
AN-Pv3	X	X	234	147	0,51	0,29	0,63	0,37
AN-Pv4	X[3]	X	259	162	-	0,37	-	0,42
AN-Pv5	X[3]	X	259	229	-	0,16	-	0,15
AN-Pv8	X	-	387	-	0,37	-	0,39	-
AN-Pv9	X[3]	X[4]	233	177	-	-	-	-
AN-Pv10	X	X	214	131	0,6	0,42	0,66	0,5
AN-Pv17	X[3]	X	115	223	-	0,15	-	0,15
AN-Pv18	X	-	409	-	0,51	-	0,63	-
AN-Pv22	X	X	340	106	0,06	**0,03**	**0,03**	**0,02**
AN-Pv26_1	X	X	205	234	**0,01**	**0**	**0,01**	**0**
AN-Pv28	X	X	88	236	0,6	0,6	0,66	0,66
AN-Pv29	X	X	79	168	0,6	0,6	0,66	0,66
AN-Pv30	X	-	216	-	0,22	-	0,17	-
AN-Pv32	X[3]	X[3]	138	146	-	-	-	-
AN-Pv33	X	-	225	-	**0,03**	-	**0,03**	-
AN-Pv35	X[3]	X	46	55	-	0,29	-	0,35
AN-Pv41	-	X	-	134	-	0,58	-	0,66
AN-Pv44	X	X[3]	344	109	0,65	-	0,73	-
AN-Pv46	X	-	381	-	0,2	-	0,15	-
AN-Pv47	X	X	439	100	0,22	0,1	0,17	0,07
AN-Pv51	X	X	319	383	0,6	0,5	0,66	0,56
AN-Pv54	X	-	399	-	0,6	-	0,66	-
AN-Pv55	X[3]	X[3]	391	212	-	-	-	-
AN-Pv57	X	X	242	217	0,7	0,98	0,75	0,99
AN-Pv63	X	-	601	-	0,29	-	0,27	-
AN-Pv64	X[3]	X	230	309	-	0,07	-	**0,05**
AN-Pv66	X	-	266	-	0,37	-	0,42	-
AN-Pv68	X	X	523	206	0,6	0,65	0,68	0,75
AN-Pv69	X	X[4]	274	129	0,06	-	**0,04**	-
gssE18	X	X	71	202	0,15	0,11	0,11	0,08
gssE20	X[3]	X[4]	81	198	-	-	-	-
AN-PvCO	X	X	539	112	0,61	0,5	0,68	0,56
AN-TGA	X	X	173	409	0,53	0,06	0,66	**0,04**

AN-DNAJ	X	-	598	-	**0,03**	-	**0,02**	-
g510	X	X	342	181	0,10	0,61	0,07	0,73
g523	X	-	363	-	**0,03**	-	**0,02**	-
Leg044	X[3]	X	103	765	-	0,47	-	0,56
Leg100	X[3]	X	48	518	-	0,29	-	0,27
Leg133	X	X	241	345	0,35	**0,02**	0,37	**0,01**
Leg223	X[3]	X	146	322	-	**0,02**	-	**0,01**
PvSHP1	X	X	97	163	0,54	0,09	0,66	**0,05**

Tabella 2. Lista di loci utilizzati, lunghezze delle porzioni analizzate e FDR associati ad ogni locus analizzato. I valori in grassetto rappresentano contig putativamente sotto selezione. [1] Le analisi di selezione sono state fatte su 39 loci, che includono regioni esoniche o introniche polimorifiche; [2] X indica la presenza/assenza di esoni o introni sequenziati all'interno di ogni locus; [3] regioni esoniche/introniche monomorfiche tra la forma domesticata e selvatica; [4] regioni esoniche/introniche monomorfiche tra la forma domesticata e selvatica dopo la rimozione di gap.

Per verificare se i geni identificati come selezionati durante la domesticazione (direttamente o indirettamente a causa di *hitchiking*) siano bersagli noti di selezione, abbiamo confrontato il nostro set di geni con i risultati di altri studi (Bellucci et al. 2014; Schmutz et al. 2014). L'allineamento con BLAT, validato con BLASTN, contro il trascrittoma oggetto del precedente studio (Bellucci et al. 2014) ha identificato una corrispondenza per 23 dei 43 loci analizzati. Due dei loci putativamente selezionati di questo studio (An-Pv69 e Leg223) corrispondono ai contig Ref_259_comp14324 and Ref_25_comp4672, due contig identificati come sottoposti a selezione durante la domesticazione del fagiolo comune in Mesoamerica anche nello studio del trascrittoma (Bellucci et al. 2014). Dei rimanenti 21 contig con una corrispondenza nel trascrittoma, 3 risultano selezionati unicamente in questo studio, mentre gli altri risultano non selezionati in entrambi gli studi.

La possibilità di mappare i 43 loci in esame nel genoma di riferimento del fagiolo comune (Schmutz et al. 2014) ha permesso un confronto con le analisi di selezione svolte da (Schmutz et al. 2014). In particolare, nello studio dove viene presentato il genoma di riferimento, è stata investigata la storia della domesticazione del fagiolo comune ri-sequenziando un campione di 160 genotipi: 30 selvatici mesoamericani, 74 domesticati mesoamericani, 30 selvatici andini e 26 domesticati andini. In questi campioni è stata calcolata la diversità genetica (π) e la statistica di differenziamento della popolazione (F_{st}) in finestre di 10 kb con uno spostamento di 2kb e in singoli

geni. Una finestra o un gene sono stati considerati selezionati se ricadevano nell'intervallo superiore al 90% della distribuzione empirica per il rapporto $\pi_{selvatico}/\pi_{domesticato}$ e per la statistica F_{st}. Otto loci sono stati individuati nelle finestre di 10 kb putativamente selezionate durante la domesticazione del pool genico andino da (Schmutz et al. 2014), dieci nelle finestre selezionate durante la domesticazione del pool genico mesoamericano. Di questi loci, rispettivamente uno (AN-Pv22) e quattro (AN-Pv26.1, AN-Pv33, Leg133 e Leg223) sono stati identificati come selezionati nel nostro studio. Il confronto con i singoli geni candidati identificati da (Schmutz et al. 2014) (dove la diversità genetica e la statistica di differenziamento della popolazione sono utilizzate su singoli geni per identificare loci differenziati a bassa variabilità nelle popolazioni) sembra più affidabile, in quanto identifica un gene preciso e non una finestra che copre una vasta regione del genoma. Sei geni utilizzati in questo studio sono stati identificati come sotto selezione durante la domesticazione da (Schmutz et al. 2014). Tre geni sono geni candidati nel Mesoamerica (Pvul.001G143100, Pvul.007G113700 e Pvul.009G231900) che corrispondono a AN-Pv33, AN-Pv66 e Leg223, rispettivamente; An-Pv33 e Leg223 sono stati identificati come putativamente selezionati anche in questo studio. Analogamente, nelle Ande, tre geni del nostro studio (AN-Pv5, AN-Pv69 e AN-DNAJ) sono stati identificati nel lavoro di Schmutz come geni candidati (Phvul.007G177200, Phvul.002G242000, Phvul.002G257300), di cui due (AN-Pv69 e AN-DNAJ) sono stati individuati anche in questo studio. Questo risultato è interessante in quanto questo studio si focalizza solamente nell'evento di domesticazione del Mesoamerica.

Discussione

In questo studio è stato utilizzato un gruppo di geni omologhi a geni noti di *Arabidopsis thaliana*, disponibili in letteratura, per valutare l'effetto della domesticazione, sia in termini di variazioni di diversità genetica, sia in termini di geni selezionati.

L'analisi della diversità genetica mostra dei livelli di diversità significativamente più bassi nel gruppo domesticato del Mesoamerica rispetto al gruppo selvatico, sia nei livelli di π, che nei livelli di θ. La riduzione di diversità nella forma domesticata, espressa con l'indice di Vigouroux, è stimata in circa il 50%, con una riduzione

maggiore nella regione non codificante. Un pattern simile è evidenziato nel confronto dei siti polimorfici condivisi tra il pool domesticato e mesoamericano, dove il numero di siti polimorfici nella popolazione selvatica monomorfici nella domestica (135) è di circa 12 volte maggiore rispetto al caso contrario (11 siti polimorfici in MD e monomorfici in MW). Questi risultati supportano il profondo effetto della domesticazione sulla variabilità genetica della popolazione domesticata del fagiolo comune, risultato già evidenziato nel precedente studio sul trascrittoma (Bellucci et al. 2014). Curiosamente la riduzione della variabilità nella popolazione domesticata e il differenziamento tra le due forme (selvatica e domesticata) è maggiore nelle regioni non codificanti, supportando l'idea che la domesticazione abbia avuto un forte impatto sull'intero genoma, sia a causa delle riduzioni demografiche associate, sia a causa del linkage fisico con regioni selezionate.

Le analisi di selezione hanno portato all'individuazione di 7 loci utilizzando il modello 1 per stimare la distribuzione nulla della statistica di selezione, 11 loci utilizzando il modello 2. I loci individuati con il modello 1 sono compresi nell'insieme di quelli individuati nel modello 2, suggerendo come i due modelli siano molto simili.

Il confronto tra il set di geni analizzati in questo studio con i geni dello studio precedente sul trascrittoma (Bellucci et al. 2014), mostra come solo 23 dei loci analizzati siano in comune. Questo potrebbe essere spiegato con il fatto che il trascrittoma dello studio precedente non copre tutta l'espressione genica della specie, ma solo quella della prima foglia trifogliata. Tra 23 i geni presenti, 2 sono condivisi e selezionati in entrambi, 3 sono selezionati solo in questo lavoro, nessuno è selezionato solo nel trascrittoma, suggerendo un buon livello di sovrapposizione tra i risultati dei due studi. L'identificazione di tre geni addizionali in questo studio potrebbe essere spiegata dal diverso modello usato per stimare la distribuzione nulla della statistica di selezione (con un tasso di mutazione diverso per le regioni introniche, una differente dimensione campionaria e una diversa parametrizzazione della lunghezza del gene nel modello), e suggerisce un ulteriore indagine per capire come modelli diversi possano influenzare l'identificazione della selezione naturale.

Il set di geni analizzati in questo studio è stato ulteriormente confrontato con i risultati delle analisi sui genomi completi svolto da (Schmutz et al. 2014). Otto dei nostri loci

cadono all'interno di finestre identificate come sotto selezione nel lavoro di (Schmutz et al. 2014) nel pool andino, dieci all'interno di finestre selezionate nel pool mesoamericano. Alcuni di questi loci sono stati identificati come target di selezione anche in questo studio, rispettivamente uno nelle finestre andine e quattro nelle finestre mesoamericane. La discrepanza tra i risultati è spiegabile dal fatto che l'approccio di (Schmutz et al. 2014) non tiene in considerazione il modello demografico sottostante, quindi le finestre considerate potrebbero essere finestre a bassa variabilità legate al profondo impatto demografico che la domesticazione ha avuto sulla specie. Inoltre l'uso di ampie finestre potrebbe non essere adeguato per l'identificazione di singoli geni target di selezione, ma solo di regioni genomiche di adattamento. I risultati della strategia a singolo locus utilizzato nel lavoro di (Schmutz et al. 2014) sono più simili ai risultati di questo studio: tra i sei geni identificati come outlier da (Schmutz et al. 2014) e condivisi con il nostro studio, quattro sono stati identificati da noi come selezionati. La strategie di (Schmutz et al. 2014) ha identificato due di questi geni come target putativi di selezione nel pool andino. Considerato che il nostro approccio è mirato sull'identificazione di selezione legata all'evento di domesticazione nel Mesoamerica, questi geni potrebbero rappresentare un evento di evoluzione convergente, dove la stessa pressione selettiva in pool geograficamente distinti porta geni comuni ad essere selezionati.

Conclusione

Questo studio ha identificato un set di loci che potrebbero essere stati rilevanti nella domesticazione del fagiolo comune in Mesoamerica, suggerendo un'ulteriore investigazione del loro ruolo in altre specie (facilitato dal fatto che sono noti ortologhi in *Arabidopsis thaliana*) e analisi funzionali che potrebbero aiutare nell'identificazione di varianti utili per la coltivazione del fagiolo.

Il confronto con studi precedenti ha confermato la forte riduzione nei livelli di variabilità associati alla domesticazione osservata in precedenza. L'analisi di selezione ha fornito risultati parzialmente sovrapponibili con studi precedenti, mostrando come approcci diversi, o versioni diverse dello stesso approccio, possano portare a risultati diversi. Questo supporta la necessità di ulteriori analisi suoi metodi,

ad esempio attraverso studi di simulazione, allo scopo di acquisire maggiori informazioni sulla potenza e sul tasso di falsi positivi di questi test.

Materiali supplementari.

Tabella S1. Parametri demografici del Modello 1.

Parametro	Descrizione	Distribuzione	Media	Dev. Std.	Min	Max
TANC	Tempo di divergenza tra il pool genico Andino e Mesoamericano	Normale	111000	40000	67330	192835
TBAB	Tempo dell'inizio del collo di bottiglia Andino	Normale	98845	3000	94051	104027
TEAB	Tempo della fine del collo di bottiglia Andino	Normale	67858	1500	64655	70865
TBMD	Tempo dell'inizio della domesticazione in Mesoamerica	Normale	8160	133	7922	8426
TBAD	Tempo dell'inizio della domesticazione nelle Ande	Normale	8500	8	8495	8517
TEMD	Tempo della fine della domesticazione in Mesoamerica	Normale	6260	150	5971	6567
TEAD	Tempo della fine della domesticazione nelle Ande	Normale	7012	35	6945	7075
NANC	Dimensione della popolazione effettiva ancestrale	Normale	418000	105000	266000	628000
NBA	Dimensione della popolazione effettiva Andina durante il collo di bottiglia	Normale	105000	20000	65000	142000
NMD	Dimensione della popolazione effettiva domesticata Mesoamericana	Uniforme			100000	100000
NMW	Dimensione della popolazione effettiva selvatica Mesoamericana	Normale	292000	240000	125000	773000
NAW	Dimensione della popolazione effettiva selvatica Andina	Normale	137000	182000	70000	502000
NAD	Dimensione della popolazione effettiva domesticata Andina	Uniforme			100000	100000
IMD	Intensità del collo di bottiglia della domesticazione in Mesoamerica (percentuale)	Normale	47,65	3	41,66	52,13
IAD	Intensità del collo di bottiglia della domesticazione nelle Ande (percentuale)	Normale	47,26	0,5	46,25	48,59
MWD	Tasso di migrazione dalla popolazione selvatica alla domesticata	Uniforme			0,000001	0,00001
XM	Fattore di migrazione asimmetrico	Uniforme			2	6
MDW	Tasso di migrazione dalla popolazione domesticata alla popolazione selvatica	XM*MWD				
MU	Tasso di mutazione (per sito per generazione,esoni)	Lognormale	1,00E-09	2,50E-09	5,00E-10	5,00E-09
MU	Tasso di mutazione (per sito per generazione,introni)	Lognormale	1,00E-08	2,50E-08	5,00E-09	5,00E-08

Tabella S2. Parametri demografici del Modello 2.

Parametro	Descrizione	Distribuzione	Media	Dev. Std.	Min	Max
NMW	Dimensione della popolazione effettiva selvatica Mesoamericana	Normale	561000	50000	463300	658300
NMD	Dimensione della popolazione effettiva domesticata Mesoamericana	Uniforme			100000	100000
NAW	Dimensione della popolazione effettiva selvatica Andina	Normale	219000	25000	188500	271300
NAD	Dimensione della popolazione effettiva domesticata Andina	Uniforme			100000	100000
MDW	Tasso di migrazione dalla popolazione domesticata alla popolazione selvatica	XM*MWD				
MWD	Tasso di migrazione dalla popolazione selvatica alla domesticata	Uniforme			0,000001	0,00001
MMWAW	Tasso di migrazione dalla popolazione selvatica Mesoamericana alla popolazione selvatica Andina	Normale	0,0000004	0,000000026	0,000000357	0,000000452
MAWMW	Tasso di migrazione dalla popolazione selvatica Andina alla popolazione selvatica Mesoamericana	Normale	0,00000026	0,000000023	0,000000214	0,000000297
TEMD	Tempo della fine della domesticazione in Mesoamerica	Normale	6260	150	5971	6567
TEAD	Tempo della fine della domesticazione nelle Ande	Normale	7012	35	6945	7075
TBMD	Tempo dell'inizio della domesticazione in Mesoamerica	Normale	8160	133	7922	8426
TBAD	Tempo dell'inizio della domesticazione nelle Ande	Normale	8500	8	8495	8517
TEAB	Tempo della fine del collo di bottiglia Andino	TANC-LAB				
TANC	Tempo di divergenza tra il pool genico Andino e Mesoamericano	Normale	165000	10000	146200	183700
MU	Tasso di mutazione (per sito per generazione,esoni)	Lognormale	1,00E-09	2,50E-09	5,00E-10	5,00E-09
MU	Tasso di mutazione (per sito per generazione,introni)	Lognormale	1,00E-08	2,50E-08	5,00E-09	5,00E-08
IMD	Intensità del collo di bottiglia della domesticazione in Mesoamerica (percentuale)	Normale	47,65	3	41,66	52,13
IAD	Intensità del collo di bottiglia della domesticazione nelle Ande (percentuale)	Normale	47,26	0,5	46,25	48,59
XM	Fattore di migrazione asimmetrico	Uniforme			2	6
NANC	Dimensione della popolazione effettiva ancestrale	Normale	168000	500	158900	176200
NMWANC	Dimensione della popolazione effettiva ancestrale Mesoamericana	Normale	155000	25000	124900	205800
NAWANC	Dimensione della popolazione effettiva ancestrale Andina	Normale	3590	2750	2304	8978
LAB	Durata del collo di bottiglia Andino	Normale	75900	12000	60370	99470

Il cambiamento della modalità riproduttiva in *Zootoca vivipara*

Una delle sfide più grandi nell'ambito della biologia evoluzionistica è quella di identificare le basi genetiche dei fenotipi che differenziano individui, popolazioni o specie. Quest'area di ricerca è molto sviluppata per alcuni organismi modello (uomo, topo), ed è oggi possibile per questi organismi svolgere un'analisi di associazione su tratti fenotipici complessi correggendo per numerose fonti di errore e supportando il risultato con esperimenti funzionali. Per gli organismi non modello la ricerca in quest'area presenta ancora molti limiti, legati all'assenza di informazioni chiare sia sul fenotipo (ad es. studi funzionali adeguati) che sul genotipo (ad es. mancanza di genoma di riferimento o informazioni sulla struttura della popolazione).

In questo capitolo della tesi si è cercato di identificare polimorfismi associati a due diverse modalità riproduttive (ovipara e vivipara) in un organismo non modello, la lucertola *Zootoca vivipara*. Per questo studio è stato utilizzato l'approccio RAD-seq per ottenere un dataset di 46.314 loci RAD contenenti 82.494 SNPs in 40 individui di popolazioni ovipare e vivipare. Dopo una prima analisi esplorativa, per valutare le relazione tra i diversi individui e le diverse popolazioni alla luce delle informazioni note in letteratura, sono stati utilizzati tre diversi approcci per identificare polimorfismi potenzialmente associati alla diversa modalità riproduttiva, uno legato al differenziamento genetico delle popolazioni e due legati direttamente all'associazione genotipo-fenotipo. Sono stati in seguito individuati i geni omologhi dei 175 loci individuati dai tre metodi nella specie più vicina disponibile nei database genomici (*Anolis carolinensis*). 37 loci sono stati assegnati a un gene di *Anolis carolinensis*, e ne è stata valutata la funzione e l'interazione con gli altri geni. 17 di questi geni sono regolatori della trascrizione o proteasi, due classi di proteine già note in letteratura per avere implicazioni nel cambiamento della modalità riproduttiva, e rappresentano dei candidati d'elezione per analisi funzionali volte a spiegare i meccanismi che mediano il cambiamento di modalità riproduttiva nelle popolazioni.

Questo lavoro è stato svolto in collaborazione con il Dr. Luca Cornetti e il Dr. Cristiano Vernesi, che si sono occupati del campionamento, della produzione del dato e delle analisi preliminari dello studio e con il gruppo del professor Michael B. Thompson dell'Università di Sidney, che ha fornito il suo contributo come esperto dei cambiamenti fisiologici legati alla diversa modalità riproduttiva per valutare la rilevanza dei geni individuati. Il mio contributo personale in questo studio è legato alla parte di analisi, dove mi sono occupato dell'associazione dei polimorfismi con le diverse modalità riproduttive (nello specifico, gli approcci con F_{st} e con il software GEMMA) e della caratterizzazione delle categorie funzionali e delle interazioni dei loci identificati.

Introduzione

Il cambiamento della modalità riproduttiva ha influenzato in maniera profonda la storia degli organismi viventi, ed è stato chiaramente associato con l'evoluzione del complesso assortimento di strutture morfologiche e funzioni fisiologiche che caratterizzano la riproduzione. Gli scienziati sono sempre stati affascinati dalla variazione della modalità riproduttiva tra i diversi regni (Holsinger 2000; Touchon and Warkentin 2008) e perfino all'interno della stessa specie (Sandrock, Schirrmeister, and Vorburger 2011), ma la comprensione dell'insieme delle variazioni genomiche associate, o che guidano questa transizione, è molto limitata.

I due tipi principali di modalità riproduttiva nei vertebrati sono l'oviparità (genitori che depongono le uova) e la viviparità (gli embrioni si sviluppano all'interno del corpo del genitore, che in seguito partorisce prole viva). La relazione tra la madre e gli zigoti, in termini di scambio di gas, acqua, trasporto dei nutrienti e immunità, è molto diversa nei gruppi vivipari e ovipari (Van Dyke, Brandley, and Thompson 2014). Il confronto di tratti specifici, geni o interi genomi tra gruppi con diverse modalità riproduttive è quindi essenziale per scoprire i meccanismi alla base dell'evoluzione della viviparità.

La viviparità rappresenta un'evoluzione del tratto oviparo avvenuta circa in 140 eventi indipendenti nei vertebrati, di cui l'80% nei rettili squamati (Sites, Reeder, and Wiens 2011). All'interno dell'ordine degli Squamata vi sono casi particolari in cui linee strettamente imparentate mostrano modalità riproduttive alternative (in altri termini, esistono popolazioni ovipare e vivipare all'interno della stessa specie). Ad esempio,

almeno tre specie mostrano popolazioni con diverse modalità riproduttive localizzate in aree geografiche distinte: gli scincidi australiani *Lerista bougainvillii* e *Saiphos equalis* (Smith, Austin, and Shine 2001; Qualls and Shine 1998) e la lucertola lacertide euroasiatica *Zootoca vivipara* (Surget-Groba et al. 2006). Queste specie rappresentano modelli ideali per studiare le modificazioni morfologiche e fisiologiche, così come i processi genetici, che hanno portato alla transizione dall'oviparità alla viviparità.

Zootoca vivipara è distribuita tra l'Europa e l'Asia, ed è uno dei rettili con l'areale di distribuzione più ampio (Figura 1). La maggior parte delle popolazioni, localizzate dal Giappone all'Europa centrale, compresa la Scandinavia e le Isole Britanniche, sono vivipare e vengono classificate come *Zootoca vivipara vivipara*. L'analisi del DNA mitocondriale suggerisce l'esistenza di clade mitocondriali distinti nella popolazione vivipara, classificati come *Zootoca vivipara pannonica* e *Zootoca vivipara sachalinensis* (Surget-Groba et al. 2001). Considerato l'obiettivo di questo capitolo verrà utilizzata la classificazione *Zootoca vivipara vivipara* per indicare l'intero gruppo delle popolazioni vivipare. Due popolazioni geograficamente distinte mostrano modalità riproduttiva ovipara e sono classificate come *Zootoca vivipara louislantzi* e *Zootoca vivipara carniolica*. La prima si trova in alcune zone dei Pirenei e, dall'analisi del DNA mitocondriale, risulta geneticamente affine al clade viviparo. Questo supporta l'ipotesi di un ritorno al tratto ancestrale oviparo da un antenato viviparo (Surget-Groba et al. 2006). La seconda è distribuita lungo la regione alpina dell'Italia settentrionale, Austria meridionale, Slovenia e Croazia, ed è oggi considerata la forma ovipara ancestrale da cui sono derivate le altre popolazioni (Surget-Groba et al. 2001). In letteratura vi sono evidenze contrastanti legate alla presenza di ibridizzazione naturale tra i gruppi vivipari e ovipari in due località delle Alpi (Cornetti et al. 2015).

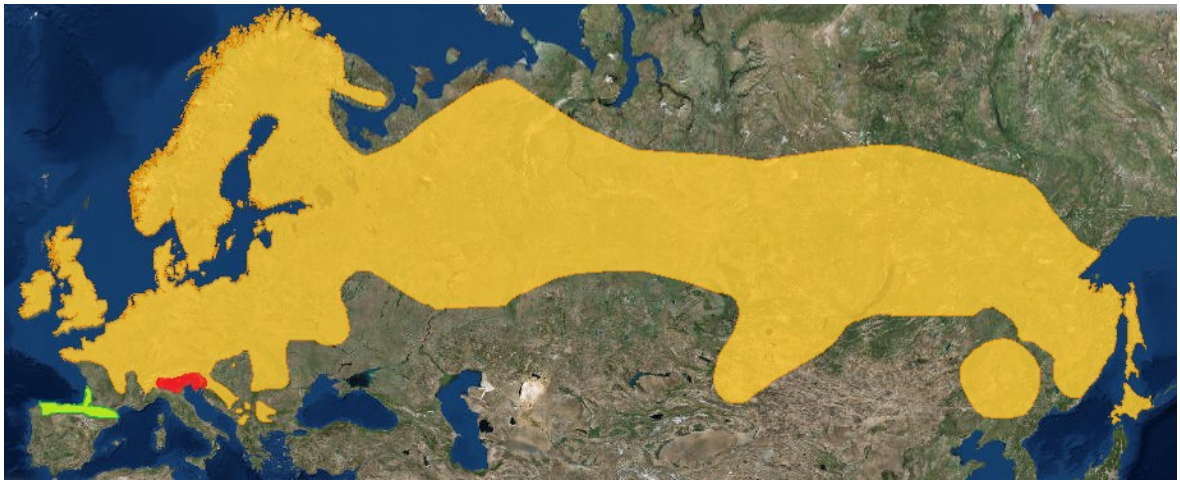

Figura 1. Areale di distribuzione della specie *Zootoca vivipara* (da IUCN 2015). In giallo è evidenziato il territorio occupato dalla sottospecie vivipara *Z. v. vivipara*, in verde l'area occupata dalla sottospecie ovipara *Z. v. louislantzi,* in rosso l'area occupata dalla sottospecie ovipara *Z. v. carniolica.*

Un aspetto importante legato alla distribuzione geografica di questa specie è relativo al fatto che le popolazioni vivipare possono essere trovate oltre il Circolo Polare Artico, mentre le popolazioni ovipare sono localizzate esclusivamente lungo il margine meridionale dell'areale di distribuzione della specie (Creemers et al. 2014), dove le temperature sono più alte. Questo scenario è in accordo con "l'ipotesi del clima freddo" (l'evoluzione della viviparità è più verosimile in climi freddi) e con l'osservazione che le specie vivipare sono più diffuse nei climi freddi rispetto alle specie che depositano le uova (Shine 2005; Rodríguez-Díaz and Braña 2012). In *Zootoca vivipara*, la viviparità si è probabilmente evoluta come conseguenza di pressioni selettive causate dal clima freddo delle fasi glaciali del Pleistocene (Surget-Groba et al. 2001): la nuova modalità riproduttiva ha in seguito permesso alle popolazioni vivipare di colonizzare la maggior parte dell'Eurasia (Cornetti et al. 2014).

L'evoluzione della viviparità richiede dei cambiamenti nello sviluppo dei tessuti per renderli in grado di supportare la gravidanza. Questi cambiamenti includono meccanismi regolativi per mantenere gli embrioni nell'utero durante lo sviluppo, meccanismi per permettere lo scambio dei gas durante la gravidanza e meccanismi per trasportare il calcio agli embrioni in assenza di calcio derivato dal guscio dell'uovo (Thompson and Speake 2006; Murphy and Thompson 2011; Stewart

2013). L'evoluzione di queste funzioni negli organismi si ottiene attraverso variazioni nella sequenza codificante dei geni, in grado alterare la funzione delle proteine, e attraverso variazioni nell'espressione genica (Rawn and Cross 2008). Queste variazioni, inducendo a livello tissutale l'espressione di geni normalmente espressi altrove nell'organismo, possono portare all'acquisizione di nuove funzioni da parte dei tessuti (True and Carroll 2002). Diversi studi hanno indagato questi aspetti in diverse specie di rettili con modalità riproduttive alternative, con risultati a tratti contrastanti. Ad esempio, studi comparativi tra le popolazioni ovipare e vivipare di *Zootoca vivipara* non hanno mostrato differenze nei livelli di espressione dei geni delle interleuchine e dei loro recettori (Paulesu et al. 2005). In aggiunta, studi nel gene angiogenico *VEGF* non hanno permesso di osservare profili di espressione correlati con la modalità riproduttiva in *Saiphos equalis* (Whittington et al. 2015). Al contrario, il sequenziamento del trascrittoma dell'utero del sincide gongillo (*Chalcides ocellatus*) e dello sincide d'erba meridionale (*Pseudemoia entrecasteauxii*) ha evidenziato cambiamenti a livello della regolazione genica durante la gravidanza (Brandley et al. 2012; Griffith 2015). Considerato che la viviparità richiede verosimilmente cambiamenti sostanziali nell'espressione genica per supportare la gravidanza, le variazioni genetiche responsabili dell'evoluzione della viviparità sono probabilmente avvenute in regioni del genoma legate alla regolazione genica.

In questo studio è stata utilizzata la tecnologia RAD-seq (*Restriction associated DNA*) su popolazioni di *Zootoca vivipara* con diversa modalità riproduttiva. I nuovi dati genomici sono stati utilizzati per comprendere meglio la struttura geografica della specie. Sono state poi ricercate le regioni genomiche o i geni associati alle diverse modalità riproduttive attraverso il confronto con le frequenze alleliche di popolazioni ovipare e vivipare e con test di associazione genotipo-genotipo. Le regioni individuate sono state associate a geni omologhi nella specie più vicina con il genoma di riferimento annotato (*Anolis carolinensis*), e sono state valutate la funzione e le interazioni di questi geni.

Materiali e metodi

Raccolta del campione, estrazione del DNA e preparazione della libreria

Quaranta estremità di coda di *Zootoca vivipara* sono state campionate in modo da coprire la maggior parte delle linee mitocondriali descritte in (Surget-Groba et al. 2006). I tessuti di coda utilizzati provenivano da individui già analizzati in altri studi di filo geografia (Surget-Groba et al. 2006; Cornetti et al. 2014; Cornetti et al. 2015). Il gene mitocondriale *citocromo b* era stato in precedenza sequenziato per tutti gli individui ed è stato utilizzato per attribuire agli individui l'appartenenza a un particolare clade, in quanto nessun tratto morfo metrico permette la classificazione dei singoli individui nelle varie sottospecie (Guillaume et al. 2006). Il campione ottenuto include 10 individui della sottospecie ovipara *Zootoca vivipara carniolica* dalle Alpi europee (mtDNA: clade A), 7 individui della sottospecie ovipara *Zootoca vivipara louislantzi* dai Pirenei (mtDNA: clade B), 17 individui della sottospecie vivipara *Zootoca vivipara vivipara* dalle Alpi europee (mtDNA: clade E) e 6 individui vivipari di *Zootoca vivipara vivipara* che rappresentano due clade mitocondriali addizionali (4 dal clade D, 2 dal clade F). Tutti i clade mitocondriali maggiori sono stati quindi considerati (Tabella 1).

Campione	Clade MtDNA	Sottospecie	Origine	Modalità riproduttiva
13_11_A	A	*Z.v.carniolica*	Italia	Ovipara
193_09_A	A	*Z.v.carniolica*	Italia	Ovipara
22_11_A	A	*Z.v.carniolica*	Italia	Ovipara
26_11_A	A	*Z.v.carniolica*	Italia	Ovipara
37_08_A	A	*Z.v.carniolica*	Italia	Ovipara
3_08_A	A	*Z.v.carniolica*	Italia	Ovipara
42_08_A	A	*Z.v.carniolica*	Italia	Ovipara
43L_A	A	*Z.v.carniolica*	Italia	Ovipara
60L_A	A	*Z.v.carniolica*	Italia	Ovipara
63L_A	A	*Z.v.carniolica*	Italia	Ovipara
10H_B2	B	*Z.v.louislantzi*	Francia	Ovipara
15H_B1	B	*Z.v.louislantzi*	Francia	Ovipara
20H_B1	B	*Z.v.louislantzi*	Francia	Ovipara
26H_B1	B	*Z.v.louislantzi*	Francia	Ovipara
35H_B2	B	*Z.v.louislantzi*	Francia	Ovipara

4H_B2	B	*Z.v.louislantzi*	Francia	Ovipara
7H_B2	B	*Z.v.louislantzi*	Francia	Ovipara
59H_D	D	*Z.v.vivipara*	Russia	Vivipara
61H_D	D	*Z.v.vivipara*	Russia	Vivipara
62H_D	D	*Z.v.vivipara*	Russia	Vivipara
65H_D	D	*Z.v.vivipara*	Romania	Vivipara
16_11_E	E	*Z.v.vivipara*	Italia	Vivipara
21_11_E	E	*Z.v.vivipara*	Italia	Vivipara
30_11_E	E	*Z.v.vivipara*	Italia	Vivipara
31_11_E	E	*Z.v.vivipara*	Italia	Vivipara
33L_E	E	*Z.v.vivipara*	Italia	Vivipara
34_11_E	E	*Z.v.vivipara*	Italia	Vivipara
35Lb_E	E	*Z.v.vivipara*	Italia	Vivipara
36Lb_E	E	*Z.v.vivipara*	Italia	Vivipara
40_11_E	E	*Z.v.vivipara*	Italia	Vivipara
42L_E	E	*Z.v.vivipara*	Italia	Vivipara
46L_E	E	*Z.v.vivipara*	Italia	Vivipara
47L_E	E	*Z.v.vivipara*	Italia	Vivipara
50L_E	E	*Z.v.vivipara*	Italia	Vivipara
63_08_E	E	*Z.v.vivipara*	Italia	Vivipara
806_09_E	E	*Z.v.vivipara*	Italia	Vivipara
80_09_E	E	*Z.v.vivipara*	Italia	Vivipara
811_09_E	E	*Z.v.vivipara*	Italia	Vivipara
79H_F	F	*Z.v.vivipara*	Austria	Vivipara
80H_F	F	*Z.v.vivipara*	Austria	Vivipara

Tabella 1. Dettaglio dei campioni analizzati nello studio, con ID individuo, clade mitocondriale, sottospecie, origine e modalità riproduttiva.

Il DNA genomico è stato estratto usando il kit QIAGEN DNeasy Blood & Tissue kit (QIAGEN Inc., Hilden, Germania). Il DNA è stato trattato con RNaseA (QIAGEN) e successivamente quantificato con il fluorimetro Qubit 2.0 (Invitrogen). Per ridurre la complessità del genoma sono state preparate diverse librerie frammentando il DNA genomico con un enzima di restrizione (RAD sequencing, (Baird et al. 2008a)). 1 µg di DNA per ogni individuo è stato digerito con l'enzima SbfI in un volume di reazione di 50 µl. L'adattatore P1, contenente un barcode unico per ogni individuo, è stato legato ad ogni campione. I campioni, marcati individualmente, sono stati raggruppati

e frammentati in frammenti di circa 500 bp con l'ultrasonicatore Covaris S220 (Covaris Inc, Woburn MA, USA). In seguito i frammenti di 300-500 bp sono stati selezionati su gel di agarosio, dimensione ideale per il sequenziamento con una flow cell Illumina. La preparazione della libreria è stata completata legando adattatori P2, permettendo così il sequenziamento dei soli frammenti con incorporati entrambi i frammenti P1 e P2. La libreria è stata sequenziata su una flow cell Illumina HiSeq2000 con protocollo Paired End presso il centro GenePool di Edimburgo, Scozia.

Genotipizzazione SNP

La qualità dalle *reads* Illumina ottenute è stata valutata con FastQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) per valutare l'impatto del calo della qualità della lettura lungo la lunghezza della sequenza. *Reads* con barcodes ambigui, siti di restrizione ambigui o di bassa qualità (con accuratezza della base chiamata < 99.99%) sono state scartate. In aggiunta, le *reads* identificate come duplicati di PCR (identiche in entrambi i paired-end, legate al sequenziamento di cloni dello stesso frammento in cluster diversi) sono state ridotte a una singola copia. Le rimanenti *reads* single-end, fiancheggianti i siti di restrizione, sono state utilizzate per la chiamata degli SNPs usando il software Stacks v 1.02 (Catchen et al. 2013). Nello specifico, lo script Perl denovo_map.pl (incluso in Stacks) è stato usato per: 1) allineare le *reads* single end a formare dei gruppi di *reads* simili impilate tra loro (stack), utilizzando la soglia minima di 6 *reads* per chiamare uno stack; 2) confrontare stacks per ottenere un set di loci e chiamare SNP usando un algoritmo di massima verosimiglianza (Hohenlohe et al. 2010a); 3) costruire un catalogo di loci contro cui confrontare tutti i campioni. Il programma *populations*, implementato in Stacks, è stato in seguito utilizzato per esportare i genotipi.

Valutazione della struttura genetica tra le popolazioni

Il dataset di polimorfismi ottenuto con Stacks è stato utilizzato per calcolare le distanze genetiche tra gli individui, distanza definita come 1 – il numero di alleli in comune sul totale. Lo scaling multidimensionale a due dimensioni è stato applicato sulla matrice delle distanze tra gli individui utilizzando la funzione *cmdscale* di R (R Core Team 2015), al fine rappresentare gli individui come punti in un piano

cartesiano e valutare le loro relazioni genetiche. Le distanze sono state inoltre utilizzate per rappresentare gli individui su un albero filogenetico con un algoritmo *neighbor joining*, utilizzando il pacchetto *ape* di R (R Core Team 2015).

Definizione degli SNPs potenzialmente implicati nel cambiamento di modalità riproduttiva

Per definire i loci potenzialmente implicati nel cambiamento della modalità riproduttiva sono stati utilizzati tre diversi metodi indipendenti in un sottoinsieme di loci con un basso livello di dati mancanti (meno del 50% di dati mancanti). Gli SNPs selezionati con almeno 2 dei 3 metodi sono stati considerati come potenzialmente legati alla transizione nella modalità riproduttiva.

Il primo metodo si basa sull'uso di outliers di F_{st}, outliers definiti come SNPs che rispettano le seguenti condizioni: 1) F_{st} >= 0.5 tra *Zootoca vivipara vivipara* (vivipara) e *Zootoca vivipara carniolica* (ovipara); 2) F_{st} >=0.5 tra *Zootoca vivipara vivipara* (vivipara) e *Zootoca vivipara louislantzi* (ovipara); 3) F_{st} <=0.05 tra *Zootoca vivipara carniolica* (ovipara) e *Zootoca vivipara louislantzi* (ovipara). L'ultima condizione assume esplicitamente che la modalità riproduttiva ovipara (sia ancestrale che derivata) sia legata agli stessi geni.

Il secondo metodo (GEMMA) permette di calcolare l'associazione statistica tra genotipo e fenotipo attraverso l'algoritmo Genome-wide Efficient Mixed Model Association (Zhou and Stephens 2012). E' stato utilizzato il Modello Univariato Lineare Misto con il test di Wald, utilizzando come file di input un sub-set dei genotipi con basso impatto dei dati mancanti e una relatedness matrix (una matrice che quantifica le relazioni tra gli individui) stimata dai genotipi. I p-value del test di associazione sono stati corretti per test multipli (Benjamini and Hochberg 1995) e gli SNPs con p-value corretti inferiori a 0,05 sono stati considerati come SNPs di loci potenzialmente associati al tratto riproduttivo.

Il terzo metodo testa l'associazione statistica tra il genotipo e la modalità riproduttiva utilizzando TASSEL 4.0 (Bradbury et al. 2007). E' stato utilizzato il Modello Lineare Misto (MLM , (Z. Zhang et al. 2010)) corretto per la struttura di popolazione con una matrice Q basata sull'output del software STRUCTURE e sulla matrice di kinship.

L'associazione del genotipo con il tratto modalità riproduttiva (codificata come 0 per l'oviparità, 1 per la viviparità) è stata considerata significativa per p-value inferiori a 0,05, dopo aver corretto per test multipli (Benjamini and Hochberg 1995).

Associazione di polimorfismi sotto selezione con geni annotati

Per identificare geni omologhi ai loci RAD-seq di *Zootoca vivipara* in *Anolis carolinensis*, i loci RAD-seq sono stati blastati contro un dataset di 21.913 geni di *Anolis carolinensis* (estratti dal database ENSEMBL usando BIOMART (Smedley et al. 2015)), con il software dc-megablast (Morgulis et al. 2008). Gli accoppiamenti individuati sono stati filtrati per lunghezza e similarità usando dei geni omologhi noti tra *Zootoca vivipara* e *Anolis carolinensis* per definire una soglia appropriata (almeno 80% di identità e 40% di lunghezza dell'allineamento). Gli omologhi identificati sono stati poi associati a una categoria Gene Ontology (GO) utilizzando l'utility web-based PantherDb (Thomas et al. 2003).

Un test di sovra-rappresentazione è stato eseguito utilizzando un tool integrato in PantherDB, confrontando le proporzioni di termini Gene Ontology nel set di loci putativamente associati alla transizione della modalità riproduttiva con l'intero set di geni del genoma di *Anolis carolinensis*. Le relazioni dirette e indirette tra i geni potenzialmente coinvolti nel cambiamento di modalità riproduttiva sono state in seguito valutate con STRING (Szklarczyk et al. 2015).

Risultati

146.439.826 *reads* paired-end sono state sequenziate a seguito della procedura di sequenziamento con protocollo RAD-seq. Queste *reads* rappresentano dati grezzi, che sono stati in seguito filtrati per eliminare diverse fonti di errore o di ambiguità. Le *reads* con barcodes ambigui (24.270), con siti di restrizione ambigui (31.474.031) e con bassa accuratezza nel base-calling (5.830.599) sono state scartate. I cloni di PCR, corrispondenti al 59,4% del totale, sono stati ridotti a singola copia. I genotipi sono stati chiamati sulle rimanenti 22.197.925 *reads* single end. Utilizzando l'informazione del decadimento della qualità fornito da FastQC, le ultime 25 bp di queste *reads* sono state tagliate per garantire una chiamata degli SNPs affidabile. Dopo la chiamata dei polimorfismi con STACKS sono stati sono stati ottenuti 46.314

contigs contenenti 82.494 SNPs, derivati da *reads* single-end con non più di 5 polimorfismi al loro interno. Questo dataset di polimorfismi è stato utilizzato per descrivere i livelli generali di variazione genetica all'interno e tra i diversi gruppi.

Lo scaling plot multidimensionale (MDS, Figura 2) supporta l'esistenza di 3 gruppi genomici distinti, che corrispondono alle diverse modalità riproduttive e alle diverse storie evolutive delle popolazioni ovipare inferite in questa specie (Surget-Groba et al. 2001). Lungo il primo asse, l'ovipara *Zootoca vivipara carniolica* (la linea ancestrale) si separa da un gruppo composto dagli individui vivipari di *Zootoca vivipara vivipara* e gli individui ovipari di *Zootoca vivipara louislantzi. Zootoca vivipara vivipara* e *Zootoca vivipara louislantzi* si separano lungo l'asse Y. La struttura genetica si può osservare anche nell'albero degli individui in *Zootoca vivipara* (Figura 3), dove gli individui appartenenti a diversi clade mitocondriali e campionati in diversi Paesi clusterizzano in gruppi distinti.

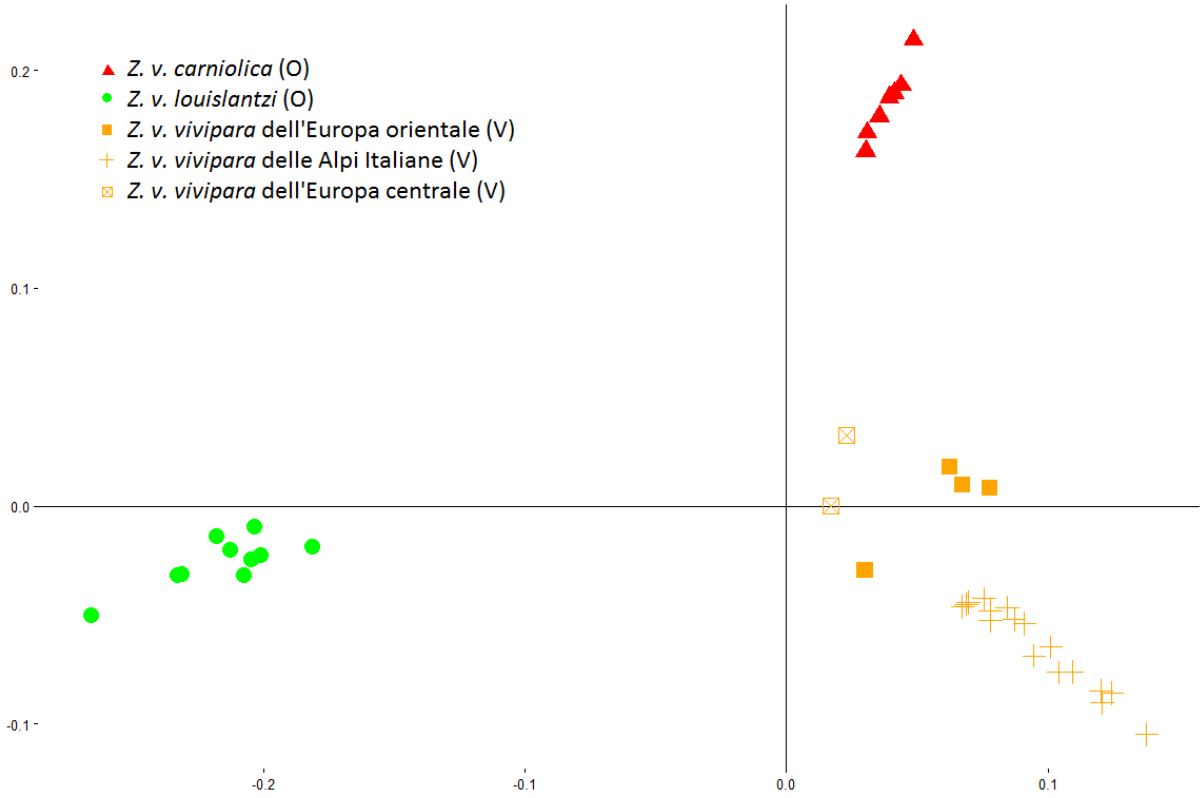

Figura 2. Scaling plot multidimensionale su 82.494 SNPs in 40 individui di *Zootoca vivipara vivipara*.

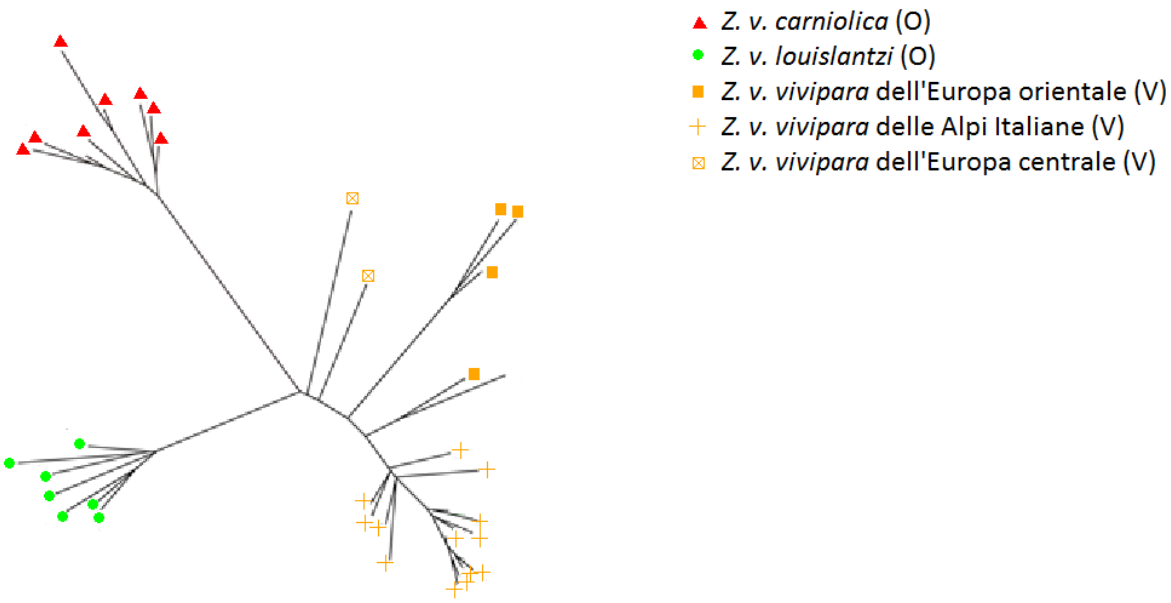

Figura 3. Albero delle distanze individuali su 82.494 SNPs in 40 individui di *Zootoca vivipara vivipara*.

L'identificazione dei geni associati alla diversa modalità riproduttiva è stata eseguita su un set ristretto di 4.908 SNPs con basso livelli di dati mancanti (meno del 50% di dati mancanti). Nella Figura 4 sono riportati gli SNPs identificati dai tre diversi approcci e il numero di SNPs identificati condivisi da più approcci. Utilizzando il criterio di loci identificati da almeno 2 su 3 approcci, 217 SNPs in 175 loci sono stati considerati come associati al cambiamento di modalità riproduttiva. 37 di questi risultano omologhi a geni di *Anolis carolinensis (*Tabella S1*)*. Il loro ruolo biologico è eterogeneo e nessuna categoria GO risulta sovra-rappresentata. Le analisi di STRING mostrano un'interazione in 9 coppie di geni con un medio livello di confidenza (0,4). Questa relazione rimane significativa per 2 coppie anche con il più alto livello di confidenza (0,9) (Tabella S2).

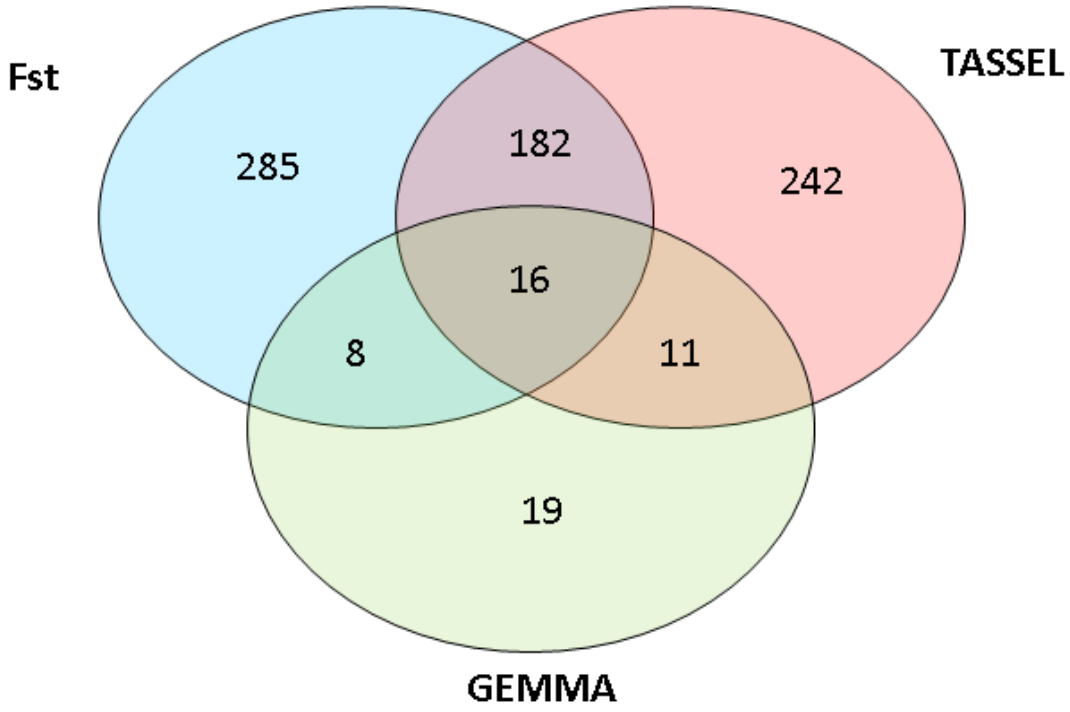

Figura 4. SNPs putativamente associati al cambiamento della modalità riproduttiva identificati dai 3 approcci.

Discussione

L'analisi esplorativa per valutare le relazioni genetiche tra gli individui di *Zootoca vivipara* in esame è in accordo con quanto noto in letteratura per marcatori mitocondriali e microsatelliti (Surget-Groba et al. 2001), e permette di individuare tre gruppi maggiori di individui corrispondente alle popolazioni/sottospecie a cui questi appartengono. La distanza genetica tra gli individui *di Zootoca vivipara carniolica e Zootoca vivipara louislantzi* rappresenta un'ulteriore conferma e contributo all'ipotesi del ritorno alla modalità ovipara da parte di un antenato viviparo in *Zootoca louislantzi* (Surget-Groba et al. 2006).

Un totale di 217 SNPs in 175 loci sono stati associati al cambiamento della modalità riproduttiva, e rappresentano regioni geniche, regolative o regioni neutrali in linkage fisico con regioni funzionali. Questo set di loci, seppur rilevante e con elevata

confidenza alla luce dei criteri utilizzati, non è onnicomprensivo, in quanto per la natura stessa della tecnologia RAD-seq è stato possibile individuare solamente i loci contenenti la sequenza dell'enzima di restrizione utilizzato. Producendo un dataset di diversa natura (ad es. sequenziamento dell'intero genoma o trascrittoma) sarà possibile ottenere un set più completo dei loci potenzialmente implicati nel cambiamento di modalità riproduttiva.

Tra i 175 loci individuati, 37 risultano omologhi a geni in *Anolis carolinensis*. I loci non associati a geni rappresentano loci in regioni neutrali o geni per i quali non è stato possibile stabilire una relazione di omologia con geni in *Anolis carolinensis* (specie più vicina disponibile nel database ENSEMBL). Con il sequenziamento di nuovi genomi di rettili, più vicini evolutivamente a *Zootoca vivipara,* sarà possibile avere un quadro più completo sulla funzione dei loci individuati.

L'annotazione funzionale nei 37 geni ha rivelato un'associazione con diversi termini GO come l'organizzazione del citoscheletro, la funzione di proteasi, i processi del sistema immunitario, i processi metabolici, il trasporto, il folding delle proteine e la regolazione della trascrizione.

Il ruolo del cambiamento dell'espressione genica nel cambiamento di modalità riproduttiva

Tredici loci associati a regolatori della trascrizione risultano contenere polimorfismi associati alla diversa modalità riproduttiva. I cambiamenti nella sequenza proteica nei regolatori della trascrizione possono cambiare sia i livelli di espressione che la tipologia di geni bersaglio (Hsia and McGinnis 2003). L'effetto fisiologico dei cambiamenti genetici nelle proteine regolative dipende da dove queste proteine sono espresse, dal modo in cui i polimorfismi alterano la sequenza amminoacidica e i bersagli con cui interagiscono questi regolatori. Di questi 13 candidati, almeno 2 sono noti regolatori dello sviluppo dell'utero negli euteri. La loro presenza nel set dei geni candidati rappresenta un supporto ulteriore al ruolo dei fattori di trascrizione nell'evoluzione della modalità riproduttiva in *Zootoca vivipara*.

Dachshund Family Transcription Factor 2 (DACH2) è un co fattore della trascrizione con un dominio di interazione proteica molto conservato. Geni knockout di DACH2 e

DACH1 nei topi femmina comportano un mancato sviluppo del tratto riproduttivo femminile (Davis et al. 2008). Nei maschi questo KO non mostra effetti simili, suggerendo come questo gene sia importante per la specializzazione del tratto riproduttivo femminile nei mammiferi. In *Drosophila*, *Dachsund*, ortologo del gene umano DACH2, facilita lo sviluppo del tratto riproduttivo maschile e femminile (Keisman and Baker 2001), suggerendo come la funzione di questo gene sia generalmente conservata tra gli animali con simmetria bilaterale. DACH2 è inoltre espresso nell'ovidotto di scincidi vivipari e ovipari.

Nei mammiferi il fattore di trascrizione SOX9 è importante per lo sviluppo del fenotipo maschile. Oltre a questo ruolo chiave per lo sviluppo del maschio, SOX9 è importante anche per lo sviluppo dell'epitelio ghiandolare nell'endometrio umano (Gonzalez 2012). SOX9 è espresso nel tessuto uterino (sia in stato di gravidanza, sia in stato non riproduttivo) negli scincidi vivipari e ovipari, e potrebbe avere un ruolo importante nello sviluppo dell'utero negli squamati (Brandley et al. 2012). Questi risultati, insieme, suggeriscono come DACH2 e SOX9 funzionino come regolatori dello sviluppo dell'utero in *Zootoca vivipara*. Variazioni nelle sequenze proteiche di questi geni potrebbero comportare cambiamenti nella regolazione genica nell'utero, fornendo quei cambiamenti regolativi necessari per lo sviluppo della viviparità.

L'evoluzione della viviparità richiede cambiamenti fisiologici a livello di tre tessuti distinti: l'utero, le ovaie e le membrane extra-embrionali. Il database Human Protein Atlas (Uhlen et al. 2015) è stato utilizzato per identificare se i regolatori della trascrizione identificati fossero espressi nei tessuti riproduttivi femminili umani: endometrio, ovario o placenta. Negli umani la placenta si forma dalla membrana corio-allantoidea, omologa a quella che forma i componenti embrionici della placenta in *Zootoca vivipara*. Tutti i 13 regolatori identificati mostrano una localizzazione in questi tessuti ad eccezione di EIF4E2. Inoltre, tutti i 13 regolatori sono espressi nell'utero dello sincide d'erba meridionale (Griffith, in review), suggerendo come questi regolatori potrebbero essere dei caratteri ancestrali espressi nel tratto riproduttivo femminile amniote. Considerato che la maggior parte di questi geni è localizzata nei tessuti riproduttivi nei mammiferi ed è espressa nell'utero degli squamati, cambiamenti a livello della sequenza proteica di questi geni potrebbero

aver alterato la loro funzione, alterando quindi la fisiologia riproduttiva di *Zootoca vivipara*.

Evoluzione delle proteine effettrici

Oltre alla funzione svolta dai geni regolatori della trascrizione nel cambiamento della modalità riproduttiva, anche altri geni identificati potrebbero essere implicati nell'evoluzione della viviparità modificando le proprietà degli enzimi nei tessuti riproduttivi. Questo studio ha associato 4 geni codificanti per proteasi con il cambiamento della modalità riproduttiva. Le proteasi giocano un ruolo importante nello sviluppo dell'utero e dei tessuti placentali durante la gravidanza nei mammiferi e nei rettili (Song, Spencer, and Bazer 2005; Song et al. 2010; Griffith 2015) e sono determinanti per la gravidanza in pesci vivipari come i cavallucci marini (Whittington et al. 2015). Nel sincide gongillo (*Chalcides ocellatus*) l'espressione di *Catepshina L* rappresenta il 5% dei geni espressi nell'utero durante la gravidanza (Brandley et al. 2012). L'associazione delle proteasi con la modalità riproduttiva e l'espressione di proteasi nell'utero degli Amnioti suggerisce come cambiamenti nelle proprietà chimiche delle proteasi potrebbero portare a cambiamenti funzionali in questi enzimi. Il ruolo di queste proteasi, espresse nell'utero e nelle membrane extra embrionali di *Zootoca vivipara*, potrebbe essere quello di facilitare il rimodellamento dell'utero durante la gravidanza.

Conclusioni

L'evoluzione della viviparità dalla oviparità richiede una serie di cambiamenti fisiologici nei tessuti riproduttivi volti a facilitare l'incubazione interna degli embrioni. I risultati di questo studio supportano l'idea che questi cambiamenti fisiologici siano corroborati da modificazioni non solo a proteine effettrici ma anche a livello di regolatori dell'espressione genica. Questo concetto è supportato da un'evidenza recente di cambiamenti regolativi su larga scala implicati nell'evoluzione della gravidanza nei mammiferi (Lynch et al. 2015).

I geni individuati in questo studio, per alcune limitazioni legate alla natura della tecnologia selezionata e alle informazioni disponibili sui database genomici, non rappresentano un insieme completo dei geni implicati nel cambiamento della

modalità riproduttiva in *Zootoca vivipara*, ma forniscono una solida base di partenza per la comprensione di questo fenomeno.

Nonostante questo studio abbia identificato una parte dei geni associati con la modalità riproduttiva in *Zootoca vivipara*, questi risultati rappresentano una correlazione più che un cambiamento genetico causale implicato nella transizione tra modalità riproduttive. Saranno quindi necessarie ulteriori analisi funzionali sui geni candidati identificati in questo studio per confermare il loro ruolo nello sviluppo della viviparità e determinare i loro meccanismi d'azione.

MATERIALI SUPPLEMENTARI

TABELLA S1. Lista di geni putativamente associate al cambiamento della modalità riproduttiva in *Zootoca vivipara* con le loro funzioni biologiche, come annotato su PantherDB utilizzando il genoma di riferimento di *Anolis carolinensis*.

ENSEMBL ID	Simbolo del gene	PANTHER Family/Subfamily	PANTHER Protein Class	PANTHER GO-Slim Molecular Function	PANTHER GO-Slim Biological Process	PANTHER GO-Slim Cellular Component
ENSACAG00000011931	LOC100563157	RNA BINDING PROTEIN FOX-1 HOMOLOG 1	RNA binding protein	RNA binding	-	-
ENSACAG00000005211	unassigned	SUBFAMILY NOT NAMED	-	-	-	-
ENSACAG00000010271	ARHGEF17	RHO GUANINE NUCLEOTIDE EXCHANGE FACTOR 17	-	-	-	-
ENSACAG00000012143	CORO6	CORONIN-6	non-motor actin binding protein	actin binding	cellular process;cytoskeleton organization	cytoskeleton;intracellular
ENSACAG00000005440	TRIP12	-	-	-	-	-

ENSACAG00000008287	NCAM1	NEURAL CELL ADHESION MOLECULE 1	immunoglobulin receptor superfamily;protein phosphatase;protein phosphatase;immunoglobulin receptor superfamily;immunoglobulin superfamily cell adhesion molecule	phosphoprotein phosphatase activity;phosphoprotein phosphatase activity;receptor activity	immune system process;induction of apoptosis;cellular protein modification process;cell cycle;cell-cell signaling;cell-cell adhesion;muscle contraction;neurological system process;ectoderm development;mesoderm development;induction of apoptosis;angiogenesis;nervous system development;muscle organ development	-
ENSACAG00000011737	LOC100554830	SUBFAMILY NOT NAMED	dehydrogenase;reductase	oxidoreductase activity	carbohydrate metabolic process	-
ENSACAG00000001018	LOC100565850	TROPOMODULIN-1	non-motor actin binding protein	cytoskeletal protein binding	cellular process;system process;cellular component morphogenesis;cell differentiation;cytoskeleton organization;cellular component biogenesis	actin cytoskeleton;cytoplasm

ENSACAG00000022140	TFE3	TRANSCRIPTION FACTOR E3	basic helix-loop-helix transcription factor	sequence-specific DNA binding transcription factor activity;sequence-specific DNA binding transcription factor activity	transcription from RNA polymerase II promoter;lipid metabolic process	-
ENSACAG00000005391	UIMC1	BRCA1-A COMPLEX SUBUNIT RAP80	-	protein binding	nitrogen compound metabolic process;DNA repair;cellular protein modification process;cellular process;response to stress;regulation of nucleobase-containing compound metabolic process;chromatin organization	protein complex;nucleus;intracellular
ENSACAG00000006650	ESPL1	SEPARIN	nucleic acid binding	nucleic acid binding	mitosis	-
ENSACAG00000014237	PELI2	E3 UBIQUITIN-PROTEIN LIGASE PELLINO HOMOLOG 2	-	-	-	-

ENSACAG00000007471	TMPRSS13	TRANSMEMBRANE PROTEASE SERINE 13	peptide hormone;receptor;serine protease;serine protease;protease inhibitor;annexin;calmodulin	serine-type peptidase activity;receptor activity;calcium ion binding;hormone activity;calmodulin binding;calcium-dependent phospholipid binding;peptidase inhibitor activity	lipid metabolic process;proteolysis;response to external stimulus;lipid transport;regulation of catalytic activity	extracellular region
ENSACAG00000011262	LOC100564211	ENDOTHELIN-CONVERTING ENZYME-LIKE 1	metalloprotease;metalloprotease	metallopeptidase activity	proteolysis;cellular process;blood circulation;regulation of vasoconstriction	-
ENSACAG00000009917	PIKFYVE	1-PHOSPHATIDYLINOSITOL 3-PHOSPHATE 5-KINASE	chaperonin	-	protein folding;protein complex assembly;protein complex biogenesis	-
ENSACAG00000010833	EIF4E2	EUKARYOTIC TRANSLATION INITIATION FACTOR 4E TYPE 2	translation initiation factor	translation initiation factor activity;translation initiation factor activity;translation initiation factor activity	translation;regulation of translation	-

ENSACAG00000008719	SOX9	TRANSCRIPTION FACTOR SOX-9	HMG box transcription factor;nucleic acid binding	sequence-specific DNA binding transcription factor activity;sequence-specific DNA binding transcription factor activity	transcription from RNA polymerase II promoter;regulation of transcription from RNA polymerase II promoter	-
ENSACAG00000008253	LOC100553347	PLECTIN	non-motor actin binding protein	structural constituent of cytoskeleton;actin binding	cellular process;cellular component morphogenesis;cellular component organization	actin cytoskeleton;intracellular
ENSACAG00000006538	CLUH	CLUSTERED MITOCHONDRIA PROTEIN HOMOLOG	translation initiation factor	translation initiation factor activity;translation initiation factor activity;translation initiation factor activity	translation;regulation of translation	-
ENSACAG00000012332	SNAI2	-	-	-	-	-
ENSACAG00000016841	TBXAS1	THROMBOXANE-A SYNTHASE	oxygenase;isomerase	oxidoreductase activity;isomerase activity	immune system process;respiratory electron transport chain;fatty acid biosynthetic process;steroid metabolic process;response to external stimulus;regulation of biological process	-

ENSACAG00000016165	SEMA3E	SEMAPHORIN-3E	membrane-bound signaling molecule	receptor binding	immune system process;cell communication;neurological system process;ectoderm development;mesoderm development;angiogenesis; nervous system development;heart development	-
ENSACAG00000013802	DACH2	DACHSHUND HOMOLOG 2	transcription factor	sequence-specific DNA binding transcription factor activity;sequence-specific DNA binding transcription factor activity	ectoderm development;nervous system development	-
ENSACAG00000014366	BACH2	TRANSCRIPTION REGULATOR PROTEIN BACH2	-	-	neurological system process;anatomical structure morphogenesis	-
ENSACAG00000005363	GRINA	PROTEIN LIFEGUARD 1	receptor	receptor activity	apoptotic process;apoptotic process;negative regulation of apoptotic process	-
ENSACAG00000017962	KAT2A	HISTONE ACETYLTRANSFERASE KAT2A	acetyltransferase;chromatin/chromatin-binding protein	acetyltransferase activity;nucleic acid binding;chromatin binding	transcription from RNA polymerase II promoter;cellular process;chromatin organization	-

ENSACAG00000006614	KIF1A	KINESIN-LIKE PROTEIN KIF1A	microtubule binding motor protein	motor activity	phosphate-containing compound metabolic process;nitrogen compound metabolic process;catabolic process;nucleobase-containing compound metabolic process;cellular component movement;transport	protein complex;cytoskeleton;intracellular
ENSACAG00000027343	LOC100567135	SUBFAMILY NOT NAMED	aspartic protease;aspartic protease	aspartic-type endopeptidase activity	proteolysis	-
ENSACAG00000006878	CYP26C1	-	-	-	-	-
ENSACAG00000010899	unassigned	VOLTAGE-DEPENDENT CALCIUM CHANNEL SUBUNIT ALPHA-2/DELTA-4	-	cation channel activity;cation transmembrane transporter activity;channel regulator activity	cellular process;cation transport;regulation of biological process	plasma membrane;integral to membrane;protein complex;cell part
ENSACAG00000024617	unassigned	C2 DOMAIN-CONTAINING PROTEIN 3	-	-	-	-
ENSACAG00000016689	RSU1	-	-	-	-	-
ENSACAG00000004762	SSUH2	PROTEIN SSUH2 HOMOLOG	-	-	-	-

ENSACAG00000002271	PLCB4	1-PHOSPHATIDYLINOSITOL 4,5-BISPHOSPHATE PHOSPHODIESTERASE BETA-4	signaling molecule;phospholipase;guanyl-nucleotide exchange factor;calcium-binding protein	phospholipase activity;calcium ion binding;receptor binding;small GTPase regulator activity;guanyl-nucleotide exchange factor activity	phospholipid metabolic process;phospholipid metabolic process;cell communication;regulation of catalytic activity	-
ENSACAG00000002423	SMARCA2	GLOBAL TRANSCRIPTION ACTIVATOR SNF2L2-RELATED	DNA helicase;helicase	helicase activity;binding	DNA repair;transcription from RNA polymerase II promoter;cellular process;regulation of transcription from RNA polymerase II promoter;chromatin organization	-
ENSACAG00000000443	AVPR1A	VASOPRESSIN V1A RECEPTOR	G-protein coupled receptor	G-protein coupled receptor activity;binding	cell communication;blood circulation;response to endogenous stimulus;regulation of vasoconstriction	plasma membrane;integral to membrane;cell part
ENSACAG00000008463	ADAMTS14	A DISINTEGRIN AND METALLOPROTEINASE WITH THROMBOSPONDIN MOTIFS 14	metalloprotease;metalloprotease;extracellular matrix glycoprotein;serine protease inhibitor	metallopeptidase activity;protein binding;serine-type endopeptidase inhibitor activity	proteolysis;cellular process;regulation of catalytic activity	extracellular region

TABELLA S2. Lista delle interazioni tra i geni utilizzando STRING

Gene 1	Gene 2	Punteggio interazione	Evidenze che suggeriscono un legame funzionale	Evidenze di azioni specifiche
NCAM1	SMARCA2	0,443	Co-Mentioned in PubMed Abstracts (homologs)	
KAT2A	SMARCA2	0,707	Experimental/Biochemical Data (homologs)	Reaction, Post-translational modification,Binding
			Association in Curated Databases (homologs)	
			Co-Mentioned in PubMed Abstracts (homologs)	
TFE3	SMARCA2	0,656	Experimental/Biochemical Data (homologs),	Binding
			Co-Mentioned in PubMed Abstracts (homologs)	
TFE3	SOX9	0,789	Experimental/Biochemical Data (homologs)	Binding, Activation, Expression
			Co-Mentioned in PubMed Abstracts (homologs)	
CYP26C1	SOX9	0,526	Co-Mentioned in PubMed Abstracts (homologs)	
SNAI2	SOX9	**0,904**	Co-Mentioned in PubMed Abstracts	
DACH2	SNAI2	0,575	Co-Mentioned in PubMed Abstracts (homologs)	
PIKFYVE	PLCB4	**0,932**	Experimental/Biochemical Data (homologs)	Binding
			Association in Curated Databases	
			Co-Mentioned in PubMed Abstracts (homologs)	
TMPRSS13	ESPL1	0,438	Co-Mentioned in PubMed Abstracts	

Il sequenziatore portatile di terza generazione MinION basato su nanopori: test per lo studio di loci MHC nel camoscio alpino

Lo sviluppo delle tecnologie di sequenziamento è ancora in corso, ed ha portato ad una nuova generazione, la cosiddetta terza generazione (approccio a molecola singola), caratterizzata da una resa maggiore, da minor tempi e costi di sequenziamento, da *reads* più lunghe e da una minore quantità di materiale necessario per il sequenziamento (Schadt, Turner, and Kasarskis 2010). Esempi di queste tecnologie sono l'SMRT di Pacific Biosciences e l'Oxford Nanopore. In questo capitolo della tesi analizzerò le performance della tecnologia Oxford Nanopore, applicandola al caso studio di un locus del Sistema Maggiore di Istocompatibilità (MHC) nel camoscio alpino.

Il Sistema Maggiore di Istocompatibilità è un complesso poligenico che costituisce la più importante componente genetica del sistema immunitario dei mammiferi (Kelley, Walter, and Trowsdale 2005). Queste proteine sono implicate nell'interazione tra l'organismo e l'ambiente (ad esempio in caso di difesa da un parassita), e rappresentano quindi un ottimo candidato per esplorare processi adattivi e per trovare varianti genetiche associate alla resistenza ai patogeni. La variabilità a questi loci, alla luce del loro ruolo, è molto alta, sia a livello di cambiamenti a singola base e inserzioni/delezioni, sia a livello della struttura del gene, con la presenza di ripetizioni o duplicati (Axtner and Sommer 2007). Il sequenziamento e il corretto assemblaggio di questa regione importante rappresenta quindi una sfida.

Questo studio si colloca all'interno del progetto di ricerca "Variabilità genetica a marcatori molecolari e a loci del sistema immunitario in un campione di camosci alpini (*Rupicapra rupicapra*) del Parco Nazionale delle Dolomiti Bellunesi", collaborazione tra l'Università di Ferrara e il Parco Nazionale delle Dolomiti Bellunesi. Negli studi precedenti associati a questo progetto è stato tipizzato l'esone 2 del locus DRB (un locus del sistema MHC), che codifica per la regione di interazione con l'antigene ed è per questo molto variabile (sia entro specie, che tra specie). Questo approccio, in cui sono state sequenziate circa 270 bp con tecnologia Sanger, è stato ampliato amplificando una regione di circa 10mila paia basi del locus DRB, attorno all'esone 2, per caratterizzare meglio la regione. L'amplicone è stato sequenziato in diversi individui con tecnologia Illumina MiSeq, ma le limitazioni di

questa tecnologia per l'assembly *de novo* di regioni complesse (Alkan, Sajjadian, and Eichler 2010) ne ha reso difficile l'assemblaggio. Per rispondere a questo problema, abbiamo partecipato alla fase di beta-testing della tecnologia Oxford Nanopore testando il sequenziatore MinION, un sequenziatore portatile dal costo ridotto che permette la lettura di porzioni del genoma fino a decine di migliaia di paia basi (Laver et al. 2015).

Lo scopo principale di questo studio è quello di: A) assemblare il locus DRB, caratterizzando la presenza di regioni ad alta complessità, usando un approccio solo-Nanopore e un approccio ibrido Nanopore-Illumina; B) valutare il tasso di errore della tecnologia Oxford Nanopore su sequenze note; C) verificare la capacità di questo approccio di ricostruire un riferimento adeguato per la genotipizzazione degli individui.

Il mio contributo a questo studio è legato principalmente all'analisi bioinformatica delle *reads* prodotte con la tecnologia Oxford Nanopore, nello specifico al sequenziamento, all'assemblaggio delle *reads*, alla pulizia, al calcolo del tasso di errore e alla chiamata dei polimorfismi.

Materiali e metodi

Campioni e preparazione amplicone

Il campione utilizzato in questa analisi è stato raccolto da Agenti del Corpo Forestale dello Stato nel Parco Nazionale delle Dolomiti Bellunesi da tessuto muscolare di carcasse di camoscio (*Rupicapra rupicapra*). La sigla di riconoscimento di questi campioni è PDB (Parco Dolomiti Bellunesi).

I campioni analizzati in questa analisi sono 6 individui del Parco. Sono in seguito elencati gli identificativi degli individui con un'indicazione dello stato allelico all'esone 2, tipizzati in precedenza: PDB60b (*1/*1), PDB70 (*1/*1), PDB44b (*1/*19), PDB47 (*1/*19), PDB61 (*19/*19), PDB66 (*19/*19).

Prima dell'amplificazione del locus sono state analizzate sequenze omologhe di DRB in altre specie (*Bos taurus* e *Ovis aries*) nella banca dati GenBank (Benson et al. 2013) alla ricerca di informazioni sul grado di conservazione degli esoni attorno all'esone 2 e sulla lunghezza del locus. I primer sono stati disegnati in regioni conservate (corrispondenti alle estremità dell'esone 1 e dell'esone 3) e la porzione del locus DRB in esame è stata amplificata con una PCR lunga, utilizzando la polimerasi *Platinum Taq DNA Polymerase*

High Fidelity (Life Technologies), che permette di amplificare frammenti di lunghezza superiore alle 10 kbp (la lunghezza attesa della porzione formata dall'introne 1, l'esone 2 e l'introne 2 sulla base del confronto con le altre specie).

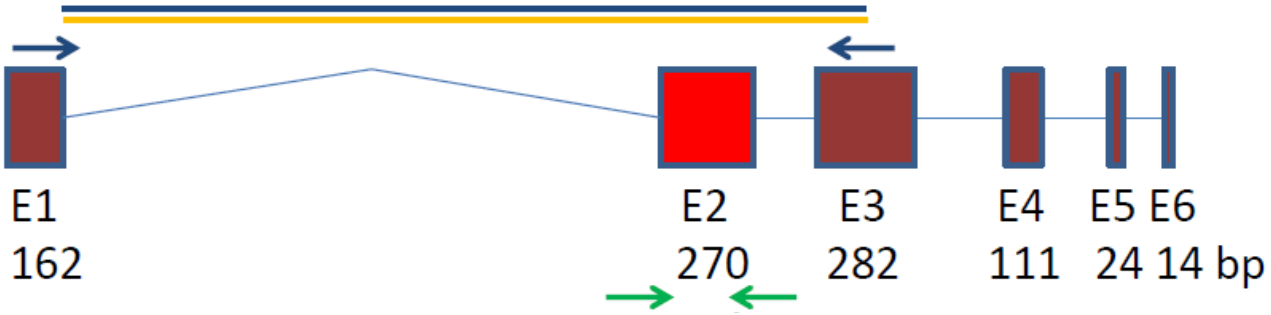

Figura 1. Schema della struttura del locus DRB1. Le frecce blu indicano i primer utilizzati per produrre l'amplicone (linee gialle e blu), le frecce rosse i primer usati per lo studio sull'esone 2 (evidenziato in rosso).

La regione amplificata è stata visualizzata su gel di agarosio, ed è stata tagliata la banda corrispondente all'amplicone per eliminare eventuali frammenti aspecifici. Il DNA è stato estratto da gel con il kit di estrazione da gel MinElute (Qiagen) e quantificato con fluorimetro Qubit BR.

Sequenziamento Illumina e assembly de novo

Il sequenziamento ad alto coverage degli ampliconi dei 6 individui è stato fatto presso il *Berlin Center for Genomics in Biodiversity Research* (Germania) su piattaforma Illumina MiSeq. Il protocollo utilizzato è l'Illumina Nextera XT DNA (Parkinson et al. 2012), un protocollo che combina la frammentazione del DNA e il legame con gli adattatori in un unico passaggio, particolarmente indicato in condizioni sperimentali con poco materiale di partenza.

Le *reads* prodotte, paired-end con lunghezza pari a 2x250bp, sono state filtrate utilizzando FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) con i seguenti criteri: 1) taglio alla fine della sequenza fino a trovare una base con qualità > 30, ed esclusione delle *reads* più corte di 200 bp; 2) esclusione delle *reads* con più del 5% delle basi con qualità inferiore a 30; 3) sostituzione della base con carattere di ambiguità (N) per le basi con qualità inferiore a 28. Il corretto appaiamento delle *reads* paired-end è stato ricostruito con uno script auto-prodotto.

Queste *reads* sono state utilizzate dal Dr. Rodrigo de Paula Baptista dell'Universidade Federal de Minas Gerais (Brasile) per l'assemblaggio *de novo* utilizzando Velvet (Zerbino and Birney 2008).

Preparazione libreria Nanopore

L'amplicone è stato preparato per il sequenziamento utilizzando il protocollo di sequenziamento SQK-MAP006, fornito dall'azienda Oxford Nanopore. Il protocollo prevede la presenza di adattatori da legare al DNA di partenza per permettere la lettura da parte della macchina della molecola, e del DNA di controllo (una porzione nota di fago lambda) per permettere all'azienda di eseguire controlli interni sull'esperimento. E' stato sequenziato unicamente l'individuo PDB70, con un unico aplotipo nell'esone 2, sia per la maggior qualità del campione, sia per ridurre l'impatto dell'alto error rate che caratterizza questa metodologia (Laver et al. 2015; Jain et al. 2015). Al termine del sequenziamento, le *reads* sono state inviate al servizio remoto di base calling Metrichor, che ha filtrato le *reads* per le quali non è stato possibile eseguire la lettura correttamente e le *reads* del DNA di controllo (*reads* di fago lambda). Le *reads* ottenute dal servizio remoto sono state scaricate e convertite dal formato HDF5 nativo della tecnologia Nanopore al formato fastq, utilizzando uno script python auto prodotto. E' stata valutata la qualità delle *reads* con FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) e la distribuzione della lunghezza. Le *reads* con lunghezza inferiore alle 1000 bp sono state eliminate, in quanto sono *reads* corte rispetto alla lunghezza attesa dell'amplicone (circa 9000 bp) che avrebbero appesantito il carico computazionale. Le *reads* sono state mappate contro la sequenza nota dell'esone 2 in PDB70 utilizzando l'allineatore di Genious (Kearse et al. 2012) (con criteri personalizzati adattatati al tipo di dato: mismatch massimi per *reads* 40%, gap massimi per *reads* 20%, minimo 25 basi identiche all'esone 2), così da dividere le *reads* di DRB da *reads* provenienti da un'eventuale contaminazione. E' stata valutata la lunghezza di questi diversi set di *reads*.

Una strategia per l'assembly *de novo* di *reads* Nanopore non è ancora disponibile, in quanto gli algoritmi esistenti sono ottimizzati per *reads* prodotto con tecnologie di seconda generazione, caratterizzate da un minor tasso di errore. Per superare questo problema ho ideato una semplice ma efficiente strategia per assemblare un contig da *reads* Nanopore. Venti *reads* di circa 9000 bp (la lunghezza attesa del locus in esame dalle analisi in laboratorio) contenenti l'esone 2 di DRB sono state estratte, e mappate contro la *reads* più

lunga di questo insieme utilizzando l'allineatore nativo di Geneious (impostando una sensitività elevata, ovvero un algoritmo computazionalmente intensivo che permette di allineare *reads* molto divergenti come nel nostro caso). Dall'allineamento è stato creato un consenso chiamando la base più frequente presente in almeno il 25% delle *reads* ed eliminando le posizioni non coperte da almeno tutte le sequenze. Questo consenso (consenso_1) rappresenta un nucleo iniziale fatto con *reads* della lunghezza nota dell'amplicone e contenenti l'esone 2. Su questo consenso sono state mappate con lo stesso algoritmo tutte le *reads* contenti l'esone 2 per raffinare la sequenza, creando in seguito un altro consenso (consenso_2) con gli stessi criteri usati in precedenza (ad eccezione del coverage minimo, alzato ad almeno 50 *reads*). Sul consenso_2 sono state infine mappate tutte le *reads* Nanopore, per includere nell'analisi anche frammenti dell'amplicone che non contengono l'esone 2, chiamando un consenso finale (consenso_3) con gli stessi criteri.

La procedura è stata ripetuta con un set di 20 *reads* iniziali diverse per valutare la stabilità della strategia.

Assemblaggio combinato Illumina-Nanopore

Due diversi approcci sono stati utilizzati per combinare i punti di forza delle due tecnologie per l'assemblaggio di uno scaffold di riferimento.

A partire dai contig individuati dalla strategia che fa uso delle *reads* Illumina, il Dr. Rodrigo de Paula Baptista ha utilizzato il software SSPACE Long-Read (Boetzer and Pirovano 2014) per riordinare i contig utilizzando le informazioni delle *reads* Nanopore, ottenendo degli scaffold.

La nostra strategia è stata quella di mappare sul consenso finale ottenuto con le *reads* Nanopore (consenso_3) le *reads* Illumina filtrate dell'individuo PDB70 (lo stesso dal quale è stata prodotta la libreria Nanopore), utilizzando l'algoritmo di allineamento di Geneious. Abbiamo scelto questo approccio per compensare l'alto tasso di errore della tecnologia Nanopore con il basso tasso di errore delle *reads* Illumina, permettendo la costruzione di un riferimento più affidabile. Il contig ottenuto è stato usato come riferimento per le analisi successive.

Le sequenze ottenute sono state confrontate tra loro utilizzando l'allineatore multiplo implementato in Geneious.

Analisi dei polimorfismi nei campioni in esame

Le *reads* Illumina filtrate dei 6 individui in esame sono state mappate sul contig ibrido Nanopore-Illumina utilizzando l'algoritmo di allineamento di Geneious. L'allineamento è stato ulteriormente raffinato rimuovendo i duplicati di PCR (copie dello stesso frammento che finiscono in diversi cluster Illumina, che potrebbero influenzare la chiamata dei polimorfismi) e riallineando attorno agli INDEL con GATK (McKenna et al. 2010) e Picard (http://broadinstitute.github.io/picard/). Da questo allineamento sono state chiamate le posizioni polimorfiche e i genotipi di tutti gli individui utilizzando il modulo HaplotypeCaller di GATK.

Stima del tasso di errore

Il tasso di errore Nanopore è stato stimato utilizzando due sequenze note disponibili per i nostri dati: la sequenza nota del DNA di controllo (porzione del fago lambda) di 3560 bp, e la sequenza nota dell'esone 2. Entrambi questi loci sono a singola copia (un solo aplotipo per l'esone 2), quindi un'eventuale differenza delle *reads* con queste sequenze rappresenta verosimilmente un errore di lettura e non un polimorfismo.

Le *reads* del DNA di controllo sono state recuperate e mappate contro la sequenza di riferimento nota del frammento di fago lambda utilizzando l'algoritmo di allineamento standard di Geneious. Le *reads* con lunghezza diversa dalla sequenza di riferimento sono state scartate. L'allineamento è stato esportato ed è stato usato uno script di R personalizzato per valutare le differenze tra il riferimento e ogni singola *reads* e stimare il tasso di errore (definito come il numero di basi differenti rispetto alla sequenza di riferimento sul totale). Sono stati contati i mismatch (basi diverse tra *reads* e referenza), le inserzioni (posizioni nella *reads* con un gap nella referenza) e le delezioni (posizioni nella *reads* con un gap rispetto alla referenza), e il tasso di errore è stato calcolato dividendo questi valori per il numero totale di confronti.

Questa strategia è stata replicata usando l'esone 2. Le *reads* Nanopore, mappate in precedenza sull'esone 2, sono state tagliate alle estremità dell'esone 2 e sono state scartate tutte le *reads* parzialmente sovrapposte. La stessa procedura utilizzata con il DNA di controllo è stata ripetuta.

Risultati

Il locus in esame è stato amplificato ed è stato estratto da gel nella banda compresa tra il marker 8 kb e il marker 10 kb. L'estrazione da gel ha rivelato la presenza di contaminazione, contaminazione persistente anche dopo successive fasi di corsa su gel ed estrazioni.

Sequenziamento ed assemblaggio Illumina

Il sequenziamento Illumina ha prodotto un totale di 1.243.596 *reads* nei 6 individui. Le *reads* sono state filtrate secondo stringenti criteri di qualità, ottenendo un dataset finale di 617.527 *reads*. Il numero di *reads* paired-end per ogni individuo, prima e dopo il filtro, è riportato in Tabella 1.

Individuo	**Raw *reads***	***Reads* dopo filtraggio**
PDB60Bis	159662	78425
PDB70	274508	140721
PDB44Bis	181947	81284
PDB47	269322	147388
PDB61	192016	102110
PDB66	166141	67599

Tabella 1. *Reads* Illumina disponibili per ogni individuo, prima e dopo l'applicazione di filtri di qualità.

Le *reads* Illumina filtrate sono state assemblate grazie al software Velvet, ottenendo un set di 16 contigs ("CONTIGS_ILLUMINA"). Questi contigs rappresentano un insieme di sequenze non sovrapposte tra di loro. Di questi contigs, uno, di lunghezza pari a 1060 bp, conteneva l'esone 2.

Sequenziamento ed assemblaggio Nanopore

Le *reads* Nanopore ottenute dall'esperimento, ottenute a partire dall'amplicone dell'individuo PDB70, sono pari a 40.585. Dopo la chiamata delle basi con Metrichor, 26.065 *reads* sono state scartate: 22.865 perché non raggiungevano la soglia di qualità standard, 3.200 perché rappresentavano *reads* del DNA di controllo. Il set di *reads* risultante, pari a 14.520, è stato ulteriormente filtrato escludendo *reads* di lunghezza inferiore alle 1000 bp, ottenendo un set finale di 12.040 *reads*.

E’ stata ricercata la presenza dell’esone 2 mappando il set finale di 12040 *reads* sulla sequenza nota dell’esone di PDB70. Circa un terzo delle *reads* (4.299) contengono l’esone 2, e rappresentano quindi un set di *reads* legate al locus DRB. L’analisi delle lunghezze delle *reads* (Figura 2) mostra come tutte le *reads* abbiano diversi picchi ma, scomponendo questo insieme in due gruppi (con esone 2, e senza esone 2), il set di *reads* contenenti l’esone 2 abbiano un picco di frequenza attorno alle 9.000 bp.

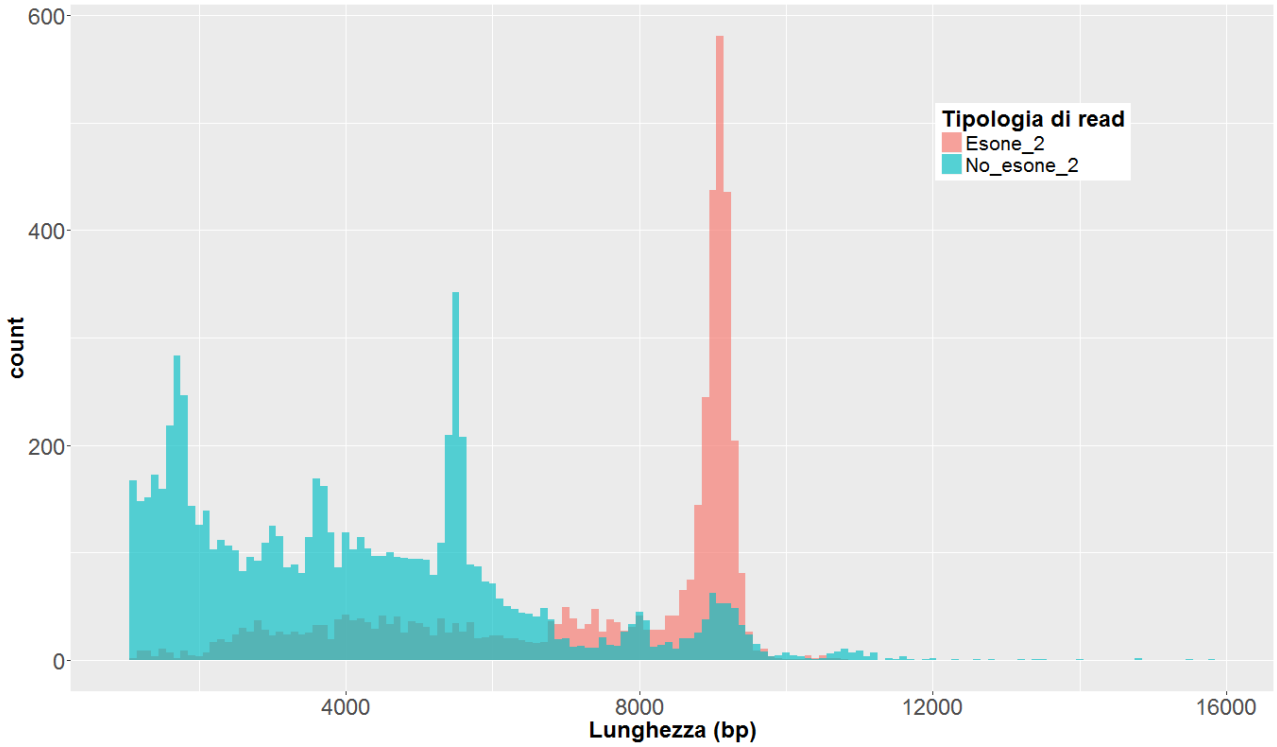

Figura 2. Distribuzione della lunghezza delle *reads* Nanopore. In rosso sono visualizzate le *reads* contenenti l’esone 2, in blu le *reads* che non contengono l’esone 2.

L’assemblaggio delle *reads* Nanopore con la procedura descritta su materiali e metodi ha portato a un consenso finale di 8829 bp. Il consenso finale contiene la sequenza nota dell’esone 2 e presenta alcune regioni ripetute. Il confronto di questo consenso con quello ottenuto da un set iniziale diverso di 20 *reads* mostra una forte identità tra le due sequenze (96.2%) e le stesse caratteristiche nella regione (esone e sequenze ripetute).

Approccio ibrido Nanopore Illumina

Il software SSPACE Long-Read ha permesso l’utilizzo delle informazioni delle letture Nanopore per riordinare i contig Illumina. Sono stati ottenuti 8 scaffold finali, uno dei quali, lungo 9669 bp, contenente l’esone 2 (“SCAFFOLD_ILL”). In questo scaffold, i nucleotidi

delle regioni non coperte dai contigs Illumina sono state sostituite con carattere di ambiguità (N).

Il mapping delle *reads* Illumina sul consenso Nanopore finale (112818 *reads* mappate su 140721, coverage medio 6219) ha permesso la chiamata di un consenso ibrido Nanopore-Illumina lungo 9052 bp ("SCAFFOLD_NAN"). Un secondo consenso ibrido Nanopore-Illumina è stato ottenuto utilizzando il consenso Nanopore ottenuto utilizzando un set iniziale di 20 *reads* diverso ("SCAFFOLD_NAN2"). I due contig ibridi ottenuti, da due consensi iniziali diversi, hanno la stessa lunghezza e sono identici, ad eccezione di 4 nucleotidi a livello di una regione micro satellite a valle dell'esone 2.

Il confronto tra i contig ottenuti con diversi approcci è presentato in Figura 3.

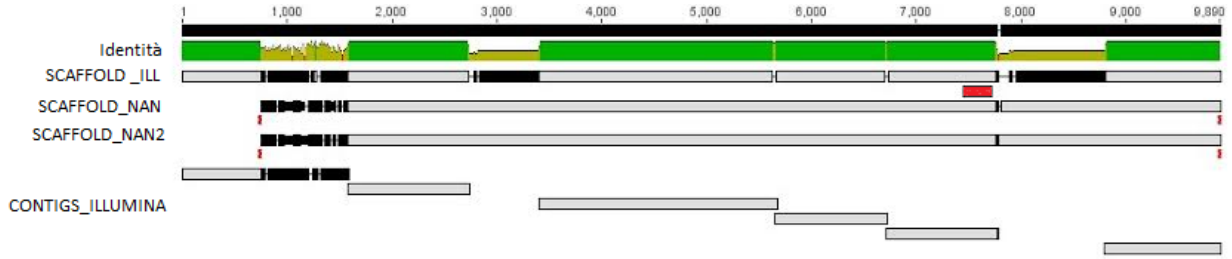

Figura 3. Confronto tra le sequenze ottenute con diversi approcci. La barra identità è verde quando tutte le sequenze confrontate sono identiche, gialla se sono diverse (indicate nelle singole sequenze dal colore nero).

Analisi dei polimorfismi nei campioni in esame

I 6 individui con stato allelico noto all'esone 2 mostrano una perfetta concordanza tra la loro sequenza nota dell'esone 2 e i genotipi chiamati nelle posizioni corrispondenti (7 siti polimorfici). 136 polimorfismi sono presenti nel resto della sequenza, alcuni dei quali rappresentano dei siti che non possono essere chiamati correttamente da Illumina. Ad esempio, 7 di questi siti mostrano uno stato con più di due alleli per inserzioni/delezioni, con una copertura molto bassa a supporto degli alleli.

Stima del tasso di errore

Le *reads* di fago lambda sono state recuperate e mappate sulla sequenza di riferimento. Le *reads* più corte o più lunghe del riferimento (artefatti del sequenziamento) sono state eliminate. Il tasso di errore della tecnologia Nanopore rispetto a una referenza di fago

lambda è stato stimato al 23,35 %, composto da 14,46% dovuto a mismatch, 5,7% dovuto a inserzioni e 3,1% dovuto a delezioni (stima su 2061 *reads*) (Figura 4).

La stima del tasso di errore utilizzando come referenza l'esone 2 mostra risultati analoghi: tasso di errore 24,27%, composto da 16,5% dovuto a mismatch, 3,8% dovuto a inserzioni e 3,8% dovuto a delezioni (stima su 4212 *reads*) (Figura 5)

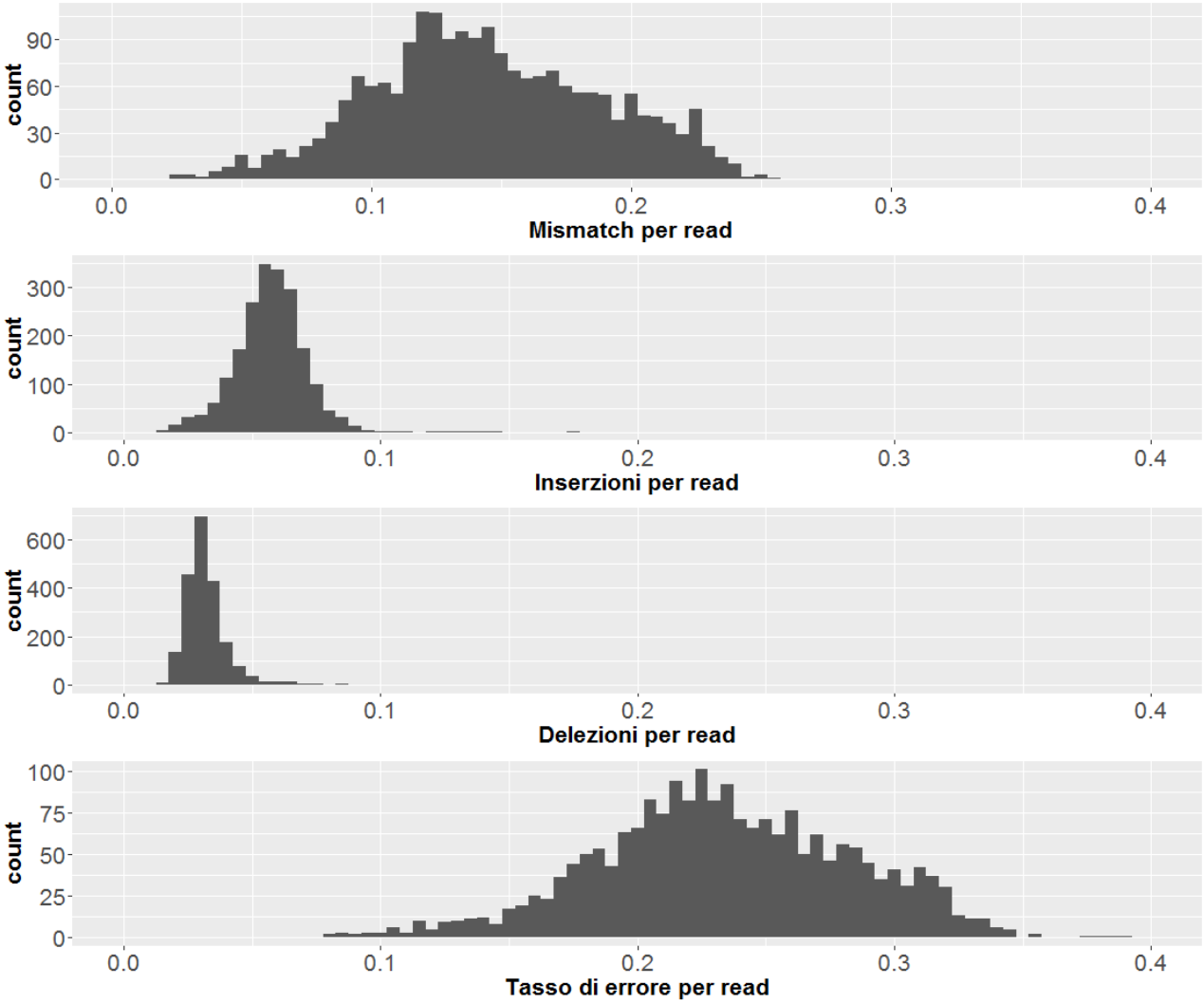

Figura 4. Distribuzione del tasso di errore su 2061 *reads* di DNA di controllo, scomposto nelle componenti mismatch, inserzioni e delezioni.

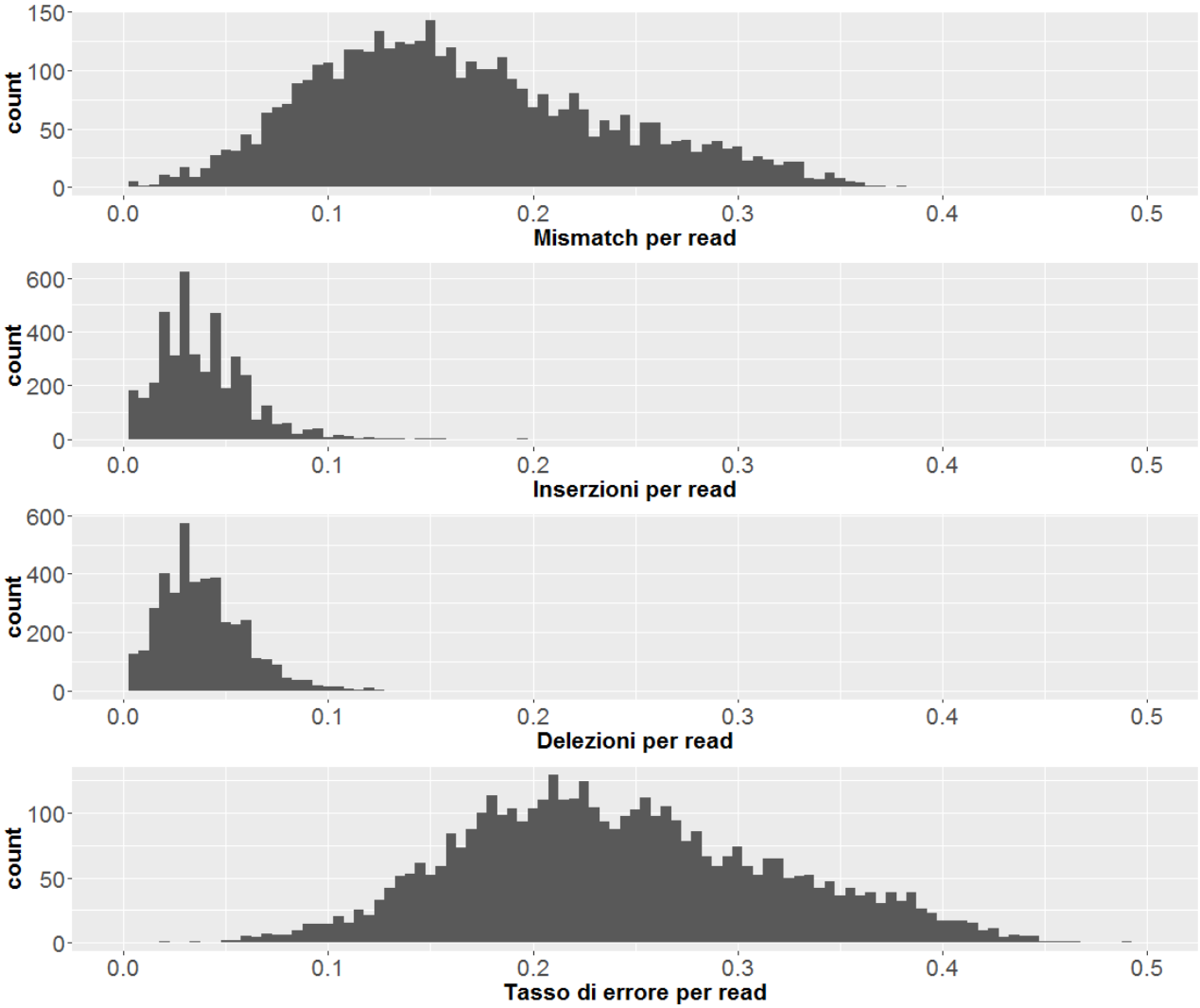

Figura 5. Distribuzione del tasso di errore sulla porzione esonica di 4212 *reads* contenenti l'esone 2, scomposto nelle componenti mismatch, inserzioni e delezioni.

Discussione

L'approccio Illumina ha generato 16 contig, uno dei quali, lungo 1060 bp contiene l'esone 2. Questo dato non corrisponde con la lunghezza attesa dell'amplicone (visualizzato nel gel in laboratorio), e rappresenta un indizio di come un approccio solo Illumina non sarebbe stato sufficiente per conoscere la sequenza dell'intero locus.

L'approccio solo Nanopore ha mostrato delle difficoltà iniziali a causa della presenza di prodotti aspecifici di PCR (presenti anche nel gel come sequenze non separabili). Il mapping dell'esone 2, con sequenza nota, ci ha permesso di separare le *reads* legate al locus DRB dal resto (lettura del locus DRB prive dell'esone 2, o un prodotto aspecifico di PCR) e di utilizzarle per chiamare un consenso. La strategia seguita per ricostruire il locus

sembra stabile (vista l'altra similarità tra due repliche della stessa procedura), e il locus ricostruito ha una lunghezza simile alla lunghezza attesa.

La stima del tasso di errore (circa 23%) è coerente con la letteratura (Laver et al. 2015; Jain et al. 2015), ed è l'ennesima conferma di come un approccio solo Nanopore potrebbe risultare limitato per definire regioni polimorfiche in presenza regioni a basso coverage.

L'informazione combinata Nanopore ed Illumina ha permesso la ricostruzione di un locus di riferimento lungo 9051 bp. Il confronto tra le sequenze ottenute con diversi approcci mostra una difficoltà generale della tecnologia Illumina nella risoluzione di regioni ripetute (i contig Illumina sono infatti interrotti a livello di queste ripetizioni). Queste difficoltà permangono anche quando i contig sono riordinati utilizzando la tecnologia Nanopore, in quanto rimangono regioni non lette o sono presenti regioni che non sembrano corrispondere a quanto letto con la tecnologia Nanopore. La ricostruzione fatta utilizzando unicamente la tecnologia Nanopore sembra più appropriata: infatti le letture Nanopore rappresentano una lettura diretta del locus in esame, mentre lo scaffold Illumina è ricostruito unendo tra loro porzioni diverse del locus. Il confronto dei singoli contig Illumina con il contig Nanopore (senza correzione Illumina) mostra però una difficoltà di quest'ultimo nel chiamare correttamente le basi in determinate posizioni: nonostante la struttura del locus sia definita, rimangono molti errori dovuti all'alto tasso di errore. Un approccio ibrido Nanopore-Illumina sembra quindi essenziale per risolvere regioni genomiche ad alta complessità: la prima tecnologia per capire la struttura e creare un riferimento grezzo, la seconda per correggere gli errori introdotti dalla prima.

Il genotipo inferito all'esone 2 è coerente con quanto osservato con il sequenziamento Sanger dell'esone 2 per tutti gli individui. Questo è una prova a sostegno che il locus DRB è a singola copia nel genoma, e non è un locus duplicato (come molti loci del sistema immunitario). Infatti, in presenza di duplicazione, avremmo osservato nuovi alleli nell'esone 2, all'interno degli stessi individui. Sono stati osservati siti polimorfici all'esterno dell'esone 2, ed è in corso il loro studio per capire se è possibile ricostruire correttamente degli aplotipi. Il campione scelto per il sequenziamento Nanopore risulta essere poco polimorfico (un solo sito eterozigote, esterno all'esone 2), e questo ci impedisce di usare la tecnologia Nanopore per la ricostruzione di un aplotipo.

Conclusioni

In questo studio abbiamo studiato l'applicabilità della tecnologia Oxford Nanopore in un locus complesso (DRB) in una specie non modello. Questo ci ha permesso di ottenere un contig di riferimento che non sarebbe stato possibile ottenere usando solamente la tecnologia Illumina, a causa della complessità del locus. Il locus ricostruito si è mostrato un'ottima referenza per il mapping delle *reads* Illumina e la chiamata dei genotipi. Nonostante sia ancora limitata a causa dell'alto tasso di errore, la tecnologia Oxford Nanopore potrebbe rivelarsi un'importante risorsa per i ricercatori, sia per le sue caratteristiche, sia per i suoi costi estremamente competitivi rispetto ai diretti concorrenti.

Considerazioni finali

Il recente sviluppo tecnologico nel settore delle indagini genomiche, e il conseguente sviluppo delle metodologie computazionali bioinformatiche e di genomica di popolazioni, sono attualmente in una fase di crescita frenetica, a volte caotica. Nuovi strumenti e nuovi metodi nascono e si estinguono prima ancora che ne vengano completamente comprese le potenzialità e le problematiche, e non esistono quindi standard condivisi e accettati unanimemente che possano guidare la scelta di un approccio, una piattaforma, o un metodo statistico a seconda della domanda scientifica a cui si è interessati. Tutto ciò si traduce spesso nella necessità di ricorrere a soluzioni *ad hoc*, la cui capacità di rispondere a problemi più generali e di diffondersi nella comunità scientifica dovrà essere valutata nell'arco dei prossimi anni.

Il mio progetto di dottorato si è sviluppato all'interno di questo contesto, fornendo un contributo per la soluzione di problematiche diverse in specifici progetti (quattro dei quali presentati in questa tesi).

Lo sviluppo del software 4P ha dimostrato come il calcolo parallelo possa ridurre significativamente i tempi computazionali quando si devono calcolare semplici statistiche di genetica di popolazioni a partire da dataset di grandi dimensioni. Questo risultato ha sottolineato l'importanza di software e algoritmi ottimizzati che facciano un uso efficiente delle risorse hardware disponibili, specialmente per l'analisi di migliaia di data set prodotti da simulazioni genomiche.

Lo studio del trascrittoma del fagiolo comune ha evidenziato l'importanza di considerare la storia demografica di una specie quando vengono studiati i pattern genetici per identificare i loci soggetti ad un processo selettivo. La storia demografica di una specie non è nota in molti casi, ma può essere stimata con un certo grado definito di incertezza. L'uso di distribuzioni a priori e diversi modelli è quindi necessario per evitare l'impatto di confondimento della demografia nella ricerca di selezione e per ridurre il tasso di falsi positivi. La seconda parte di questo studio ha evidenziato uno dei problemi maggiori dell'RNA-seq, ossia la specificità tessuto specifica che può portare a dati mancanti, e la necessità di un confronto tra diversi approcci per identificare i loci target di selezione.

L'associazione di geni con il cambiamento della modalità riproduttiva in *Zootoca vivipara* ha mostrato come differenti approcci da diversi campi di studio (studi di associazione e

analisi di differenziamento delle popolazioni) possano essere integrati per identificare geni candidati. In questo studio sono emersi alcuni problemi legati alla perdita di informazione associata con un dataset di tipo RAD-seq e l'uso di un genoma di riferimento molto divergente.

Infine, la ricostruzione di una regione del sistema maggiore di istocompatibilità nel genoma del camoscio alpino ha mostrato i limiti e le potenzialità di alcune tecnologie di seconda e terza generazione, evidenziato la necessità di un approccio ibrido per la risoluzione di regioni complesse del genoma.

Bibliografia

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. doi:10.1038/nature15393.

Akey, Joshua M., Ge Zhang, Kun Zhang, Li Jin, and Mark D. Shriver. 2002. “Interrogating a High-Density SNP Map for Signitures of Natural Selection.” *Genome Research*. doi:10.1101/gr.631202.

Alkan, Can, Saba Sajjadian, and Evan E Eichler. 2010. “Limitations of next-Generation Genome Sequence Assembly.” *Nature Methods* 8 (1): 61–65. doi:10.1038/nmeth.1527.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research*. doi:10.1093/nar/25.17.3389.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10): R106. doi:10.1186/gb-2010-11-10-r106.

Axtner, J, and S Sommer. 2007. “Gene Duplication, Allelic Diversity, Selection Processes and Adaptive Value of MHC Class II DRB Genes of the Bank Vole, Clethrionomys Glareolus.” *Immunogenetics* 59 (5): 417–26. doi:10.1007/s00251-007-0205-y.

Baird, Nathan A, Paul D Etter, Tressa S Atwood, Mark C Currey, Anthony L Shiver, Zachary A Lewis, Eric U Selker, William A Cresko, and Eric A Johnson. 2008a. “Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers.” *PloS One* 3 (10). Public Library of Science: e3376. doi:10.1371/journal.pone.0003376.

Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis, Eric U. Selker, William A. Cresko, and Eric A. Johnson. 2008b. “Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers.” *PLoS ONE* 3 (10). doi:10.1371/journal.pone.0003376.

Bellucci, Elisa, Elena Bitocchi, Alberto Ferrarini, Andrea Benazzo, Eleonora Biagetti,

Sebastian Klie, Andrea Minio, et al. 2014. “Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean.” *The Plant Cell* 26 (5). American Society of Plant Biologists: 1901–12. http://www.scopus.com/inward/record.url?eid=2-s2.0-84903627161&partnerID=tZOtx3y1.

Benazzo, Andrea, Alex Panziera, and Giorgio Bertorelle. 2015. “4P: Fast Computing of Population Genetics Statistics from Large DNA Polymorphism Panels.” *Ecology and Evolution* 5 (1). John Wiley and Sons Ltd: 172–75.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)*. doi:10.2307/2346101.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. “GenBank.” *Nucleic Acids Research* 41 (D1). doi:10.1093/nar/gks1195.

Bitocchi, Elena, Elisa Bellucci, Alessandro Giardini, Domenico Rau, Monica Rodriguez, Eleonora Biagetti, Rodolfo Santilocchi, et al. 2013. “Molecular Analysis of the Parallel Domestication of the Common Bean (Phaseolus Vulgaris) in Mesoamerica and the Andes.” *New Phytologist* 197 (1): 300–313. doi:10.1111/j.1469-8137.2012.04377.x.

Bitocchi, Elena, Laura Nanni, Elisa Bellucci, Monica Rossi, Alessandro Giardini, Pierluigi Spagnoletti Zeuli, Giuseppina Logozzo, et al. 2012. “Mesoamerican Origin of the Common Bean (Phaseolus Vulgaris L.) Is Revealed by Sequence Data.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (14): E788–96. doi:10.1073/pnas.1108973109.

Blackman, Benjamin K., David A. Rasmussen, Jared L. Strasburg, Andrew R. Raduski, John M. Burke, Steven J. Knapp, Scott D. Michaels, and Loren H. Rieseberg. 2011. “Contributions of Flowering Time Genes to Sunflower Domestication and Improvement.” *Genetics* 187 (1): 271–87. doi:10.1534/genetics.110.121327.

Boetzer, Marten, and Walter Pirovano. 2014. “SSPACE-LongRead: Scaffolding Bacterial Draft Genomes Using Long Read Sequence Information.” *BMC Bioinformatics* 15 (1): 211. doi:10.1186/1471-2105-15-211.

Bradbury, Peter J., Zhiwu Zhang, Dallas E. Kroon, Terry M. Casstevens, Yogesh Ramdoss, and Edward S. Buckler. 2007. “TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples.” *Bioinformatics* 23 (19): 2633–35. doi:10.1093/bioinformatics/btm308.

Brandley, Matthew C., Rebecca L. Young, Dan L. Warren, Michael B. Thompson, and G??nter P. Wagner. 2012. “Uterine Gene Expression in the Live-Bearing Lizard, Chalcides Ocellatus, Reveals Convergence of Squamate Reptile and Mammalian Pregnancy Mechanisms.” *Genome Biology and Evolution* 4 (3): 394–411. doi:10.1093/gbe/evs013.

Buermans, H P J, and J T den Dunnen. 2014. “Next Generation Sequencing Technology: Advances and Applications.” *Biochimica et Biophysica Acta* 1842 (10): 1932–41. doi:10.1016/j.bbadis.2014.06.015.

Catchen, Julian, Paul A. Hohenlohe, Susan Bassham, Angel Amores, and William A. Cresko. 2013. “Stacks: An Analysis Tool Set for Population Genomics.” *Molecular Ecology* 22 (11): 3124–40. doi:10.1111/mec.12354.

Chapman, Mark a, Catherine H Pashley, Jessica Wenzler, John Hvala, Shunxue Tang, Steven J Knapp, and John M Burke. 2008. “A Genomic Scan for Selection Reveals Candidates for Genes Involved in the Evolution of Cultivated Sunflower (Helianthus Annuus).” *The Plant Cell* 20: 2931–45. doi:10.1105/tpc.108.059808.

Cong, Bin, Luz S Barrero, and Steven D Tanksley. 2008. “Regulatory Change in YABBY-like Transcription Factor Led to Evolution of Extreme Fruit Size during Tomato Domestication.” *Nature Genetics* 40 (6): 800–804. doi:10.1038/ng.144.

Cornetti, L., G F Ficetola, S Hoban, and C Vernesi. 2015. “Genetic and Ecological Data Reveal Species Boundaries between Viviparous and Oviparous Lizard Lineages.” *Heredity* 115 (6): 517–26. doi:10.1038/hdy.2015.54.

Cornetti, L., Michele Menegon, Giovanni Giovine, Benoit Heulin, and Cristiano Vernesi. 2014. “Mitochondrial and Nuclear DNA Survey of Zootoca Vivipara across the Eastern Italian Alps: Evolutionary Relationships, Historical Demography and Conservation Implications.” *PLoS ONE* 9 (1). doi:10.1371/journal.pone.0085912.

Creemers, Raymond, Mathieu Denoël, João Campos, Miguel Vences, Pierre-Andre

Crochet, João Gonçalves, Philip de Pous, et al. 2014. “Updated Distribution and Biogeography of Amphibians and Reptiles of Europe.” *Amphibia-Reptilia* 35 (1): 1–31. doi:10.1163/15685381-00002935.

Crisci, Jessica L., Yu Ping Poh, Angela Bean, Alfred Simkin, and Jeffrey D. Jensen. 2012. “Recent Progress in Polymorphism-Based Population Genetic Inference.” *Journal of Heredity*. doi:10.1093/jhered/esr128.

Darwin, Charles. 1859. “On the Origins of Species by Means of Natural Selection.” *London: Murray*, 247. doi:10.1126/science.146.3640.51-b.

———. 1868. “The Variation of Animals and Plants under Domestication.” *Animals* 1: 1–411. doi:10.1017/CBO9780511709500.

Davey, John W, John L Davey, Mark L Blaxter, and Mark W Blaxter. 2010. “RADSeq: Next-Generation Population Genetics.” *Briefings in Functional Genomics* 9 (5-6): 416–23. doi:10.1093/bfgp/elq031.

Davey, John W., Paul A. Hohenlohe, Paul D. Etter, Jason Q. Boone, Julian M. Catchen, and Mark L. Blaxter. 2011. “Genome-Wide Genetic Marker Discovery and Genotyping Using next-Generation Sequencing.” *Nature Reviews Genetics* 12 (7): 499–510. doi:10.1038/nrg3012.

Davis, Richard J., Mark Harding, Yalda Moayedi, and Graeme Mardon. 2008. “Mouse Dach1 and Dach2 Are Redundantly Required for M??llerian Duct Development.” *Genesis* 46 (4): 205–13. doi:10.1002/dvg.20385.

De Alencar Figueiredo, L. F., C. Calatayud, C. Dupuits, C. Billot, J. F. Rami, D. Brunel, X. Perrier, B. Courtois, M. Deu, and J. C. Glaszmann. 2008. “Phylogeographic Evidence of Crop Neodiversity in Sorghum.” *Genetics* 179 (2): 997–1008. doi:10.1534/genetics.108.087312.

Desiderio, F, E Bitocchi, E Bellucci, D Rau, M Rodriguez, G Attene, R Papa, and L Nanni. 2012. “Chloroplast Microsatellite Diversity in Phaseolus Vulgaris.” *Frontiers in Plant Science* 3 (January): 312. doi:10.3389/fpls.2012.00312.

Dyer, W. T. Thiselton. 1877. “The Effects of Cross- and Self-Fertilisation in the Vegetable Kingdom.” *Nature* 15 (381): 329–32. doi:10.1038/015329a0.

Edgar, Robert C. 2004. “MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity.” *BMC Bioinformatics* 5: 113. doi:10.1186/1471-2105-5-113.

Eisenstein, Michael. 2012. “Oxford Nanopore Announcement Sets Sequencing Sector Abuzz.” *Nature Biotechnology* 30 (4). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 295–96. doi:10.1038/nbt0412-295.

Ewing, B, and P Green. 1998. “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities.” *Genome Research* 8 (3): 186–94. http://www.ncbi.nlm.nih.gov/pubmed/9521922.

Ewing, B, L Hillier, M C Wendl, and P Green. 1998. “Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment.” *Genome Research* 8 (3): 175–85. http://www.ncbi.nlm.nih.gov/pubmed/9521921.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. “Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data.” *Genetics* 131 (2): 479–91. doi:10.1007/s00424-009-0730-7.

Excoffier, Laurent, Isabelle Dupanloup, Emilia Huerta-Sanchez, Vitor C. Sousa, and Matthieu Foll. 2013. “Robust Demographic Inference from Genomic and SNP Data.” *PLoS Genetics* 9 (10). doi:10.1371/journal.pgen.1003905.

Excoffier, Laurent, and Matthieu Foll. 2011. “Fastsimcoal: A Continuous-Time Coalescent Simulator of Genomic Diversity under Arbitrarily Complex Evolutionary Scenarios.” *Bioinformatics* 27 (9): 1332–34. doi:10.1093/bioinformatics/btr124.

Excoffier, Laurent, and Heidi E L Lischer. 2010. “Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows.” *Molecular Ecology Resources* 10 (3): 564–67. doi:10.1111/j.1755-0998.2010.02847.x.

Fay, Justin C., and Chung I. Wu. 2000. “Hitchhiking under Positive Darwinian Selection.” *Genetics* 155 (3): 1405–13.

Foll, Matthieu, and Oscar Gaggiotti. 2008. “A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian

Perspective.” *Genetics* 180 (2): 977–93. doi:10.1534/genetics.108.092221.

Fu, Y. X., and W. H. Li. 1993. “Statistical Tests of Neutrality of Mutations.” *Genetics* 133 (3): 693–709. doi:evolution.

Gepts, P. 2002. “A Comparison between Crop Domestication, Classical Plant Breeding, and Genetic Engineering.” *Crop Science*. doi:10.2135/cropsci2002.1780.

Gepts, Paul, and Roberto Papa. 2002. “Evolution during Domestication.” *Encyclopedia of Life Sciences*, 1–7. doi:10.1038/npg.els.0003071.

Glèmin, Sylvain, and Thomas Bataillon. 2009. “A Comparative View of the Evolution of Grasses under Domestication: Tansley Review.” *New Phytologist*. doi:10.1111/j.1469-8137.2009.02884.x.

Glenn, Travis C. 2011. “Field Guide to next-Generation DNA Sequencers.” *Molecular Ecology Resources* 11 (5): 759–69. doi:10.1111/j.1755-0998.2011.03024.x.

Gonzalez, Gabriel. 2012. “ROLE OF SOX9 IN UTERINE GLAND DEVELOPMENT AND DISEASE INITIATION.” *UT GSBS Dissertations and Theses (Open Access)*. http://digitalcommons.library.tmc.edu/utgsbs_dissertations/254.

Goretti, D., E. Bitocchi, E. Bellucci, M. Rodriguez, D. Rau, T. Gioia, G. Attene, P. McClean, L. Nanni, and R. Papa. 2014. “Development of Single Nucleotide Polymorphisms in Phaseolus Vulgaris and Related Phaseolus Spp.” *Molecular Breeding* 33 (3): 531–44. doi:10.1007/s11032-013-9970-5.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52. doi:10.1038/nbt.1883.

Griffith, Oliver William. 2015. “Mechanisms of Placental Evolution: The Genetics and Physiology of Pregnancy in Lizards.” Faculty of Science. http://ses.library.usyd.edu.au:80/handle/2123/13600.

Guillaume, Cl. P., B Heulin, I Y Pavlinov, D V Semenov, a Bea, N Vogrin, and Y. Surget-Groba. 2006. “Morphological Variations in the Common Lizard, Lacerta (Zootoca) Vivipara.” *Russian Journal of Herpetology* 13 (1): 1–10.

Gutenkunst, Ryan N., Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. 2009. “Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data.” *PLoS Genetics* 5 (10). doi:10.1371/journal.pgen.1000695.

Hall, Ta. 1999. “BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT.” *Nucleic Acids Symposium Series*. doi:citeulike-article-id:691774.

Hansen, Kasper D., Steven E. Brenner, and Sandrine Dudoit. 2010. “Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming.” *Nucleic Acids Research* 38 (12). doi:10.1093/nar/gkq224.

Hatem, Ayat, Doruk Bozdağ, and Ümit V. Çatalyürek. 2011. “Benchmarking Short Sequence Mapping Tools.” In *Proceedings - 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2011*, 109–13. doi:10.1109/BIBM.2011.83.

Hecht, Valérie, Claire L Knowles, Jacqueline K Vander Schoor, Lim Chee Liew, Sarah E Jones, Misty J M Lambert, and James L Weller. 2007. “Pea LATE BLOOMER1 Is a GIGANTEA Ortholog with Roles in Photoperiodic Flowering, Deetiolation, and Transcriptional Regulation of Circadian Clock Gene Homologs.” *Plant Physiology* 144 (2): 648–61. doi:10.1104/pp.107.096818.

Hedrick, Philip W. 2005. “A Standardized Genetic Differentiation Measure.” *Evolution* 59 (8): 1633–38. doi:DOI 10.1111/j.0014-3820.2005.tb01814.x.

Hohenlohe, Paul A, Susan Bassham, Paul D Etter, Nicholas Stiffler, Eric A Johnson, and William A Cresko. 2010a. “Population Genomics of Parallel Adaptation in Threespine Stickleback Using Sequenced RAD Tags.” *PLoS Genetics* 6 (2). Public Library of Science: e1000862. doi:10.1371/journal.pgen.1000862.

Hohenlohe, Paul A., Susan Bassham, Paul D. Etter, Nicholas Stiffler, Eric A. Johnson, and William A. Cresko. 2010b. “Population Genomics of Parallel Adaptation in Threespine Stickleback Using Sequenced RAD Tags.” *PLoS Genetics* 6 (2). doi:10.1371/journal.pgen.1000862.

Hohenlohe, Paul A., Mitch D. Day, Stephen J. Amish, Michael R. Miller, Nick Kamps-Hughes, Matthew C. Boyer, Clint C. Muhlfeld, Fred W. Allendorf, Eric A. Johnson, and

Gordon Luikart. 2013. “Genomic Patterns of Introgression in Rainbow and Westslope Cutthroat Trout Illuminated by Overlapping Paired-End RAD Sequencing.” *Molecular Ecology* 22 (11): 3002–13. doi:10.1111/mec.12239.

Holsinger, Kent E. 2000. “Reproductive Systems and Evolution in Vascular Plants.” *Proceedings of the National Academy of Sciences* 97 (13): 7037–42. doi:10.1073/pnas.97.13.7037.

Hougaard, Birgit Kristine, Lene Heegaard Madsen, Niels Sandal, Marcio De Carvalho Moretzsohn, Jakob Fredslund, Leif Schauser, Anna Marie Nielsen, et al. 2008. “Legume Anchor Markers Link Syntenic Regions between Phaseolus Vulgaris, Lotus Japonicus, Medicago Truncatula and Arachis.” *Genetics* 179 (4): 2299–2312. doi:10.1534/genetics.108.090084.

Hsia, Cheryl C., and William McGinnis. 2003. “Evolution of Transcription Factor Function.” *Current Opinion in Genetics and Development*. doi:10.1016/S0959-437X(03)00017-0.

Huang, Xuehui, Nori Kurata, Xinghua Wei, Zi-Xuan Wang, Ahong Wang, Qiang Zhao, Yan Zhao, et al. 2012. “A Map of Rice Genome Variation Reveals the Origin of Cultivated Rice.” *Nature* 490 (7421): 497–501. doi:10.1038/nature11532.

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. “A Test of Neutral Molecular Evolution Based on Nucleotide Data.” *Genetics* 116 (1): 153–59. doi:10.1093/molbev/msv035.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. “Estimation of Levels of Gene Flow from DNA Sequence Data.” *Genetics* 132 (2): 583–89. doi:PMC1205159.

Hufford, Matthew B, Xun Xu, Joost van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A Cartwright, Robert J Elshire, et al. 2012. “Comparative Population Genomics of Maize Domestication and Improvement.” *Nature Genetics* 44 (7). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 808–11. doi:10.1038/ng.2309.

International Human Genome Sequencing Consortium. 2004. “International Human Genome Sequencing Consortium. Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45. doi:nature03001 [pii]\r10.1038/nature03001.

Jain, Miten, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. 2015. “Improved Data Analysis for the MinION Nanopore Sequencer.” *Nature*

Methods 12 (4). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 351–56. doi:10.1038/nmeth.3290.

Jombart, Thibaut, and Ismaïl Ahmed. 2011. “Adegenet 1.3-1: New Tools for the Analysis of Genome-Wide SNP Data.” *Bioinformatics (Oxford, England)* 27 (21): 3070–71. doi:10.1093/bioinformatics/btr521.

Jost, Lou. 2008. “GST and Its Relatives Do Not Measure Differentiation.” *Molecular Ecology* 17 (18): 4015–26. doi:10.1111/j.1365-294X.2008.03887.x.

Katsnelson, Alla. 2010. “DNA Sequencing for the Masses.” *Nature*, December. Nature Publishing Group. doi:10.1038/news.2010.674.

Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, et al. 2012. “Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data.” *Bioinformatics* 28 (12): 1647–49. doi:10.1093/bioinformatics/bts199.

Keisman, E L, and B S Baker. 2001. “The Drosophila Sex Determination Hierarchy Modulates Wingless and Decapentaplegic Signaling to Deploy Dachshund Sex-Specifically in the Genital Imaginal Disc.” *Development (Cambridge, England)* 128 (9): 1643–56. http://www.ncbi.nlm.nih.gov/pubmed/11290302.

Keller, I., C. E. Wagner, L. Greuter, S. Mwaiko, O. M. Selz, A. Sivasundar, S. Wittwer, and O. Seehausen. 2013. “Population Genomic Signatures of Divergent Adaptation, Gene Flow and Hybrid Speciation in the Rapid Radiation of Lake Victoria Cichlid Fishes.” *Molecular Ecology* 22 (11): 2848–63. doi:10.1111/mec.12083.

Kelley, James, Lutz Walter, and John Trowsdale. 2005. “Comparative Genomics of Major Histocompatibility Complexes.” *Immunogenetics*. doi:10.1007/s00251-004-0717-7.

Kent, W. James. 2002. “BLAT - The BLAST-like Alignment Tool.” *Genome Research* 12 (4): 656–64. doi:10.1101/gr.229202. Article published online before March 2002.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.” *Genome Biology* 14 (4): R36. doi:10.1186/gb-2013-14-4-r36.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. “VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing.” *Genome Research* 22 (3): 568–76. doi:10.1101/gr.129684.111.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. “ANGSD: Analysis of Next Generation Sequencing Data.” *BMC Bioinformatics* 15 (1): 356. doi:10.1186/s12859-014-0356-4.

Lam, Hon-Ming, Xun Xu, Xin Liu, Wenbin Chen, Guohua Yang, Fuk-Ling Wong, Man-Wah Li, et al. 2010. “Resequencing of 31 Wild and Cultivated Soybean Genomes Identifies Patterns of Genetic Diversity and Selection.” *Nature Genetics* 42 (12): 1053–59. doi:10.1038/ng.715.

Laver, T., J. Harrison, P.A. O’Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D.J. Studholme. 2015. “Assessing the Performance of the Oxford Nanopore Technologies MinION.” *Biomolecular Detection and Quantification* 3 (March): 1–8. doi:10.1016/j.bdq.2015.02.001.

Lester, Richard N. 1989. “Evolution under Domestication Involving Disturbance of Genic Balance.” *Euphytica* 44 (1-2): 125–32. doi:10.1007/BF00022606.

Lewontin, R. C., and J. Krakauer. 1973. “Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms.” *Genetics* 74 (1): 175–95.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

———. 2011. “Inference of Human Population History from Individual Whole-Genome Sequences.” *Nature* 475 (7357): 493–96. doi:10.1038/nature10231.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and

Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13): 1658–59. doi:10.1093/bioinformatics/btl158.

Li, WH, CI Wu, and CC Luo. 1985. "A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes." *Mol. Biol. Evol.* 2 (2): 150–74. http://mbe.oxfordjournals.org/content/2/2/150?ijkey=8c7f2fb8380b5e71827cc52fd2ed8fd1afb071ff&keytype2=tf_ipsecsha.

Li, Yun, Wei Chen, Eric Yi Liu, and Yi Hui Zhou. 2013. "Single Nucleotide Polymorphism (SNP) Detection and Genotype Calling from Massively Parallel Sequencing (MPS) Data." *Statistics in Biosciences* 5 (1): 3–25. doi:10.1007/s12561-012-9067-4.

Lin, Zhongwei, Xianran Li, Laura M Shannon, Cheng-Ting Yeh, Ming L Wang, Guihua Bai, Zhao Peng, et al. 2012. "Parallel Domestication of the Shattering1 Genes in Cereals." *Nature Genetics* 44 (6): 720–24. doi:10.1038/ng.2281.

Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. 2012. "Comparison of next-Generation Sequencing Systems." *Journal of Biomedicine and Biotechnology*. doi:10.1155/2012/251364.

Liu, Shiping, Eline D. Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou, et al. 2014. "Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears." *Cell* 157 (4): 785–94. doi:10.1016/j.cell.2014.03.054.

Lu, Jian, Tian Tang, Hua Tang, Jianzi Huang, Suhua Shi, and Chung I. Wu. 2006. "The Accumulation of Deleterious Mutations in Rice Genomes: A Hypothesis on the Cost of Domestication." *Trends in Genetics*. doi:10.1016/j.tig.2006.01.004.

Lynch, Vincent J., Mauris C. Nnamani, Aur??lie Kapusta, Kathryn Brayer, Silvia L. Plaza, Erik C. Mazur, Deena Emera, et al. 2015. "Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy." *Cell Reports* 10 (4): 551–62. doi:10.1016/j.celrep.2014.12.052.

Mamidi, S, M Rossi, S M Moghaddam, D Annam, R Lee, R Papa, and P E McClean. 2013. "Demographic Factors Shaped Diversity in the Two Gene Pools of Wild Common

Bean Phaseolus Vulgaris L." *Heredity* 110 (3): 267–76. doi:10.1038/hdy.2012.82.

Mamidi, S., Monica Rossi, Deepti Annam, Samira Moghaddam, Rian Lee, Roberto Papa, and Phillip McClean. 2011. "Investigation of the Domestication of Common Bean (Phaseolus Vulgaris) Using Multilocus Sequence Data." *Functional Plant Biology* 38 (12): 953–67. doi:10.1071/FP11124.

Mardis, Elaine R. 2008. "Next-Generation DNA Sequencing Methods." *Annual Review of Genomics and Human Genetics* 9: 387–402. doi:10.1146/annurev.genom.9.081307.164359.

Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa a Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057): 376–80. doi:10.1038/nature03959.

Maynard Smith, John, and John Haigh. 1974. "The Hitch-Hiking Effect of a Favourable Gene." *Genetical Research* 23 (1): 23–35. doi:10.1017/S0016672308009579.

McConnell, Melody, Sujan Mamidi, Rian Lee, Shireen Chikara, Monica Rossi, Roberto Papa, and Phillip McClean. 2010. "Syntenic Relationships among Legumes Revealed Using a Gene-Based Genetic Linkage Map of Common Bean (Phaseolus Vulgaris L.)." *Theoretical and Applied Genetics* 121 (6): 1103–16. doi:10.1007/s00122-010-1375-9.

McDonald, J H, and M Kreitman. 1991. "Adaptive Protein Evolution at the Adh Locus in Drosophila." *Nature* 351 (6328): 652–54. doi:10.1038/351652a0.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303. doi:10.1101/gr.107524.110.

Morgulis, Aleksandr, George Coulouris, Yan Raytselis, Thomas L. Madden, Richa Agarwala, and Alejandro A. Sch??ffer. 2008. "Database Indexing for Production MegaBLAST Searches." In *Bioinformatics*, 24:1757–64. doi:10.1093/bioinformatics/btn322.

Murphy, Bridget F., and Michael B. Thompson. 2011. "A Review of the Evolution of Viviparity in Squamate Reptiles: The Past, Present and Future Role of Molecular

Biology and Genomics." *Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology*. doi:10.1007/s00360-011-0584-0.

Nanni, L., E. Bitocchi, E. Bellucci, M. Rossi, D. Rau, G. Attene, P. Gepts, and R. Papa. 2011. "Nucleotide Diversity of a Genomic Sequence Similar to SHATTERPROOF (PvSHP1) in Domesticated and Wild Common Bean (Phaseolus Vulgaris L.)." *Theoretical and Applied Genetics* 123 (8): 1341–57. doi:10.1007/s00122-011-1671-z.

Nei, M. 1973. "Analysis of Gene Diversity in Subdivided Populations." *Proceedings of the National Academy of Sciences of the United States of America* 70 (12): 3321–23. doi:10.1073/pnas.70.12.3321.

Nei, M, and R K Chesser. 1983. "Estimation of Fixation Indices and Gene Diversities." *Annals of Human Genetics* 47 (Pt 3): 253–59. doi:10.1111/j.1469-1809.1983.tb00993.x.

Nei, M. 1978. "Estimation of Average Heterozigosity and Genetic Distance from a Small Number of Individuals." *Genetics* 89 (3): 583–90. http://www.genetics.org/content/89/3/583.abstract.

Oleksyk, Taras K, Michael W Smith, and Stephen J O'Brien. 2010. "Genome-Wide Scans for Footprints of Natural Selection." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1537): 185–205. doi:10.1098/rstb.2009.0219.

Oliver, Shannon L., Andrew J. Lenards, Roger A. Barthelson, Nirav Merchant, and Sheldon J. McKay. 2013. "Using the Iplant Collaborative Discovery Environment." *Current Protocols in Bioinformatics*, no. SUPPL.42. doi:10.1002/0471250953.bi0122s42.

Olson, M V. 1999. "When Less Is More: Gene Loss as an Engine of Evolutionary Change." *American Journal of Human Genetics* 64 (1): 18–23. doi:10.1086/302219.

Osakabe, Yuriko, Naoko Arinaga, Taishi Umezawa, Shogo Katsura, Keita Nagamachi, Hidenori Tanaka, Haruka Ohiraki, et al. 2013. "Osmotic Stress Responses and Plant Growth Controlled by Potassium Transporters in Arabidopsis." *Plant Cell* 25 (2): 609–24. doi:10.1105/tpc.112.105700.

Papa, Roberto, Elisa Bellucci, Monica Rossi, Stefano Leonardi, Domenico Rau, Paul

Gepts, Laura Nanni, and Giovanna Attene. 2007. “Tagging the Signatures of Domestication in Common Bean (Phaseolus Vulgaris) by Means of Pooled DNA Samples.” *Annals of Botany* 100 (5): 1039–51. doi:10.1093/aob/mcm151.

Parkinson, Nicholas J, Siarhei Maslau, Ben Ferneyhough, Gang Zhang, Lorna Gregory, David Buck, Jiannis Ragoussis, Chris P Ponting, and Michael D Fischer. 2012. “Preparation of High-Quality next-Generation Sequencing Libraries from Picogram Quantities of Target DNA.” *Genome Research* 22 (1): 125–33. doi:10.1101/gr.124016.111.

Paulesu, Luana, Elisa Bigliardi, Eugenic Paccagnini, Francesca Ietta, Chiara Cateni, Claude Pierre Guillaume, and Benoit Heulin. 2005. “Cytokines in the Oviparity/viviparity Transition: Evidence of the Interleukin-1 System in a Species with Reproductive Bimodality, the Lizard Lacerta Vivipara.” *Evolution and Development* 7 (4): 282–88. doi:10.1111/j.1525-142X.2005.05034.x.

Pfaffelhuber, P., A. Lehnert, and W. Stephan. 2008. “Linkage Disequilibrium under Genetic Hitchhiking in Finite Populations.” *Genetics* 179 (1): 527–37. doi:10.1534/genetics.107.081497.

Pfeifer, Bastian, Ulrich Wittelsb??rger, Sebastian E. Ramos-Onsins, and Martin J. Lercher. 2014. “PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R.” *Molecular Biology and Evolution* 31 (7): 1929–36. doi:10.1093/molbev/msu136.

Pritchard, Jonathan K., Joseph K. Pickrell, and Graham Coop. 2010. “The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation.” *Current Biology*. doi:10.1016/j.cub.2009.11.055.

Przeworski, Molly. 2002. “The Signature of Positive Selection at Randomly Chosen Loci.” *Genetics* 160 (3): 1179–89.

Purcell, S, B Neale, K Todd-Brown, L Thomas, M A R Ferreira, D Bender, J Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75. doi:10.1086/519795.

Qualls, C. P., and R. Shine. 1998. “Lerista Bougainvillii, a Case Study for the Evolution of Viviparity in Reptiles.” *Journal of Evolutionary Biology* 11 (1): 63–78.

doi:10.1007/s000360050066.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*. Vol. 1. doi:10.1007/978-3-540-74686-7.

Rawn, S M, and J C Cross. 2008. “The Evolution, Regulation, and Function of Placenta-Specific Genes.” *Annu Rev Cell Dev Biol* 24: 159–81. doi:10.1146/annurev.cellbio.24.110707.175418.

Rodríguez-Díaz, Tania, and F. Braña. 2012. “Altitudinal Variation in Egg Retention and Rates of Embryonic Development in Oviparous Zootoca Vivipara Fits Predictions from the Cold-Climate Model on the Evolution of Viviparity.” *Journal of Evolutionary Biology* 25 (9): 1877–87. doi:10.1111/j.1420-9101.2012.02575.x.

Rossi, Monica, Elena Bitocchi, Elisa Bellucci, Laura Nanni, Domenico Rau, Giovanna Attene, and Roberto Papa. 2009. “Linkage Disequilibrium and Population Structure in Wild and Domesticated Populations of Phaseolus Vulgaris L.” *Evolutionary Applications* 2 (4): 504–22. doi:10.1111/j.1752-4571.2009.00082.x.

Sabeti, P C, S F Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, T S Mikkelsen, D Altshuler, and E S Lander. 2006. “Positive Natural Selection in the Human Lineage.” *Science (New York, N.Y.)* 312 (5780). American Association for the Advancement of Science: 1614–20. doi:10.1126/science.1124309.

Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, et al. 2002. “Detecting Recent Positive Selection in the Human Genome from Haplotype Structure.” *Nature* 419 (October): 832–37. doi:10.1038/nature01027.1.

Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, et al. 2007. “Genome-Wide Detection and Characterization of Positive Selection in Human Populations.” *Nature* 449 (7164): 913–18. doi:10.1038/nature06250.

Sandrock, Christoph, Bettina E Schirrmeister, and Christoph Vorburger. 2011. “Evolution of Reproductive Mode Variation and Host Associations in a Sexual-Asexual Complex of Aphid Parasitoids.” *BMC Evolutionary Biology* 11 (1): 348. doi:10.1186/1471-2148-11-348.

Sanger, F, S Nicklen, and a R Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67. doi:10.1073/pnas.74.12.5463.

Schadt, Eric E., Steve Turner, and Andrew Kasarskis. 2010. “A Window into Third-Generation Sequencing.” *Human Molecular Genetics* 19 (R2). doi:10.1093/hmg/ddq416.

Scheet, Paul, and Matthew Stephens. 2006. “A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase.” *American Journal of Human Genetics* 78 (4): 629–44. doi:10.1086/502802.

Schiffels, Stephan, and Richard Durbin. 2014. “Inferring Human Population Size and Separation History from Multiple Genome Sequences.” *Nature Genetics* 46 (8): 919–25. doi:10.1038/ng.3015.

Schmutz, Jeremy, Phillip E McClean, Sujan Mamidi, G Albert Wu, Steven B Cannon, Jane Grimwood, Jerry Jenkins, et al. 2014. “A Reference Genome for Common Bean and Genome-Wide Analysis of Dual Domestications.” *Nature Genetics* 46 (7): 707–13. doi:10.1038/ng.3008.

Schraiber, Joshua G., and Joshua M. Akey. 2015. “Methods and Models for Unravelling Human Evolutionary History.” *Nature Reviews Genetics* 16 (12): 727–40. doi:10.1038/nrg4005.

Shine, Richard. 2005. “Life-History Evolution in Reptiles.” *Annual Review of Ecology, Evolution, and Systematics* 36 (1): 23–46. doi:10.1146/annurev.ecolsys.36.102003.152631.

Shriver, Mark D, Giulia C Kennedy, Esteban J Parra, Heather a Lawson, Vibhor Sonpar, Jing Huang, Joshua M Akey, and Keith W Jones. 2004. “The Genomic Distribution of Population Substructure in Four Populations Using 8,525 Autosomal SNPs.” *Human Genomics* 1 (4): 274–86. doi:doi:10.1186/1479-7364-1-4-274.

Sites, Jack W, Tod W Reeder, and John J Wiens. 2011. “Phylogenetic Insights on Evolutionary Novelties in Lizards and Snakes: Sex, Birth, Bodies, Niches, and Venom.” *Annual Review of Ecology, Evolution, and Systematics* 42 (1): 227–44.

doi:10.1146/annurev-ecolsys-102710-145051.

Smedley, Damian, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, et al. 2015. “The BioMart Community Portal: An Innovative Alternative to Large, Centralized Data Repositories.” *Nucleic Acids Research* 43 (W1): W589–98. doi:10.1093/nar/gkv350.

Smith, Sarah a, Christopher C Austin, and Richard Shine. 2001. “A Phylogenetic Analysis of Variation in Reproductive Mode within an Australian Lizard (Saiphos Equalis, Scincidae).” *Biological Journal of the Linnean Society* 74 (2001): 131–39. doi:10.1006/bijl.2001.0563.

Song, Gwonhwa, Daniel W Bailey, Kathrin A Dunlap, Robert C Burghardt, Thomas E Spencer, Fuller W Bazer, and Greg A Johnson. 2010. “Cathepsin B, Cathepsin L, and Cystatin C in the Porcine Uterus and Placenta: Potential Roles in Endometrial/placental Remodeling and in Fluid-Phase Transport of Proteins Secreted by Uterine Epithelia across Placental Areolae.” *Biology of Reproduction* 82 (5): 854–64. doi:10.1095/biolreprod.109.080929.

Song, Gwonhwa, Thomas E. Spencer, and Fuller W. Bazer. 2005. “Cathepsins in the Ovine Uterus: Regulation by Pregnancy, Progesterone, and Interferon Tau.” *Endocrinology* 146 (11): 4825–33. doi:10.1210/en.2005-0768.

Stewart, James R. 2013. “Fetal Nutrition in Lecithotrophic Squamate Reptiles: Toward a Comprehensive Model for Evolution of Viviparity and Placentation.” *Journal of Morphology*. doi:10.1002/jmor.20141.

Surget-Groba, Y, B Heulin, C P Guillaume, R S Thorpe, L Kupriyanova, N Vogrin, R Maslak, et al. 2001. “Intraspecific Phylogeography of Lacerta Vivipara and the Evolution of Viviparity.” *Molecular Phylogenetics and Evolution* 18 (3): 449–59. doi:10.1006/mpev.2000.0896.

Surget-Groba, Y., BENOIT HEULIN, CLAUDE-PIERRE GUILLAUME, MIKLOS PUKY, DMITRY SEMENOV, VALENTINA ORLOVA, LARISSA KUPRIYANOVA, IOAN GHIRA, and BENEDIK SMAJDA. 2006. “Multiple Origins of Viviparity, or Reversal from Viviparity to Oviparity? The European Common Lizard (Zootoca Vivipara, Lacertidae) and the Evolution of Parity.” *Biological Journal of the Linnean Society* 87

(1): 1–11. doi:10.1111/j.1095-8312.2006.00552.x.

Swanson-Wagner, R., R. Briskine, R. Schaefer, M. B. Hufford, J. Ross-Ibarra, C. L. Myers, P. Tiffin, and N. M. Springer. 2012. “Reshaping of the Maize Transcriptome by Domestication.” *Proceedings of the National Academy of Sciences* 109: 11878–83. doi:10.1073/pnas.1201961109.

Szklarczyk, Damian, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, et al. 2015. “STRING v10: Protein-Protein Interaction Networks, Integrated over the Tree of Life.” *Nucleic Acids Research* 43 (D1): D447–52. doi:10.1093/nar/gku1003.

Tajima, F. 1983. “Evolutionary Relationship of DNA Sequences in Finite Populations.” *Genetics* 105 (2): 437–60. doi:6628982.

———. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123 (3): 585–95. doi:PMC1203831.

Takahashi, Yasuyuki, and Ko Shimamoto. 2011. “Heading Date 1 (Hd1), an Ortholog of Arabidopsis CONSTANS, Is a Possible Target of Human Selection during Domestication to Diversify Flowering Times of Cultivated Rice.” *Genes & Genetic Systems* 86: 175–82. doi:10.1266/ggs.86.175.

Tassi, Francesca, Silvia Ghirotto, Massimo Mezzavilla, Sibelle Torres Vilaça, Lisa De Santi, and Guido Barbujani. 2015. “Early Modern Human Dispersal from Africa: Genomic Evidence for Multiple Waves of Migration.” *Investigative Genetics* 6: 13. doi:10.1186/s13323-015-0030-2.

The 1000 Genomes Project Consortium. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422): 56–65. doi:10.1038/nature11632.

Thomas, Paul D., Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. 2003. “PANTHER: A Library of Protein Families and Subfamilies Indexed by Function.” *Genome Research* 13 (9): 2129–41. doi:10.1101/gr.772403.

Thompson, Michael B, and Brian K Speake. 2006. “A Review of the Evolution of Viviparity in Lizards: Structure, Function and Physiology of the Placenta.” *Journal of Comparative Physiology. B, Biochemical, Systemic, and Environmental Physiology*

176 (3): 179–89. doi:10.1007/s00360-005-0048-5.

Touchon, Justin Charles, and Karen Michelle Warkentin. 2008. “Reproductive Mode Plasticity: Aquatic and Terrestrial Oviposition in a Treefrog.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (21): 7495–99. doi:10.1073/pnas.0711579105.

True, John R, and Sean B Carroll. 2002. “Gene Co-Option in Physiological and Morphological Evolution.” *Annual Review of Cell and Developmental Biology* 18: 53–80. doi:10.1146/annurev.cellbio.18.020402.140619.

Uhlen, M., L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, et al. 2015. “Tissue-Based Map of the Human Proteome.” *Science* 347 (6220): 1260419–1260419. doi:10.1126/science.1260419.

Valouev, Anton, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade, Heather Peckham, Kathy Zeng, et al. 2008. “A High-Resolution, Nucleosome Position Map of C. Elegans Reveals a Lack of Universal Sequence-Dictated Positioning.” *Genome Research* 18 (7): 1051–63. doi:10.1101/gr.076463.108.

Van Dijk, Erwin L., H??l??ne Auger, Yan Jaszczyszyn, and Claude Thermes. 2014. “Ten Years of next-Generation Sequencing Technology.” *Trends in Genetics*. doi:10.1016/j.tig.2014.07.001.

Van Dyke, James U., Matthew C. Brandley, and Michael B. Thompson. 2014. “The Evolution of Viviparity: Molecular and Genomic Data from Squamate Reptiles Advance Understanding of Live Birth in Amniotes.” *Reproduction*. doi:10.1530/REP-13-0309.

Vigouroux, Y, M McMullen, C T Hittinger, K Houchins, L Schulz, S Kresovich, Y Matsuoka, and J Doebley. 2002. “Identifying Genes of Agronomic Importance in Maize by Screening Microsatellites for Evidence of Selection during Domestication.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (15): 9650–55. doi:10.1073/pnas.112324299.

Vitti, J J, S R Grossman, and P C Sabeti. 2013. “Detecting Natural Selection in Genomic Data.” *Annu Rev Genet* 47: 97–120. doi:10.1146/annurev-genet-111212-133526.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool

for Transcriptomics." *Nature Reviews. Genetics* 10 (1): 57–63. doi:10.1038/nrg2484.

Watterson, G A. 1975. "On the Number of Segregating Sites in Genetical Models without Recombination." *Theoretical Population Biology* 7 (2): 256–76. http://www.ncbi.nlm.nih.gov/pubmed/1145509.

Watterson, G. A., and H. A. Guess. 1977. "Is the Most Frequent Allele the Oldest?" *Theoretical Population Biology* 11 (2): 141–60. doi:10.1016/0040-5809(77)90023-5.

Wegmann, Daniel, Christoph Leuenberger, Samuel Neuenschwander, and Laurent Excoffier. 2010. "ABCtoolbox: A Versatile Toolkit for Approximate Bayesian Computations." *BMC Bioinformatics* 11 (1). BioMed Central: 116. doi:10.1186/1471-2105-11-116.

Weir, B S, and C Clark Cockerham. 1984. "Estimating F-Statistics for the Analysis of Population Structure." *Evolution* 38 (6): 1358–70. doi:10.2307/2408641.

Whittington, Camilla M., Georges E. Grau, Christopher R. Murphy, and Michael B. Thompson. 2015. "Unusual Angiogenic Factor Plays a Role in Lizard Pregnancy but Is Not Unique to Viviparity." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 324 (2): 152–58. doi:10.1002/jez.b.22615.

Wright, Sewall. 1949. "The Genetical Structure of Populations." *Annals of Eugenics* 15 (1): 323–54. doi:10.1111/j.1469-1809.1949.tb02451.x.

Wright, Stephen I, Irie Vroh Bi, Steve G Schroeder, Masanori Yamasaki, John F Doebley, Michael D McMullen, and Brandon S Gaut. 2005. "The Effects of Artificial Selection on the Maize Genome." *Science (New York, N.Y.)* 308 (5726): 1310–14. doi:10.1126/science.1107891.

Wu, W, X M Zheng, G Lu, Z Zhong, H Gao, L Chen, C Wu, et al. 2013. *Association of Functional Nucleotide Polymorphisms at DTH2 with the Northward Expansion of Rice Cultivation in Asia*. *Proc Natl Acad Sci U S A*. Vol. 110. doi:10.1073/pnas.1213962110.

Xu, Xun, Xin Liu, Song Ge, JD Jeffrey D JD Jeffrey D JD Jensen, Fengyi Hu, Xin Li, Yang Dong, et al. 2012. "Resequencing 50 Accessions of Cultivated and Wild Rice Yields Markers for Identifying Agronomically Important Genes." *Nature Biotechnology* 30 (1): 105–11. doi:10.1038/nbt.2050.

Zerbino, Daniel R., and Ewan Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” *Genome Research* 18 (5): 821–29. doi:10.1101/gr.074492.107.

Zhang, Wenyu, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, and Bairong Shen. 2011. “A Practical Comparison of De Novo Genome Assembly Software Tools for next-Generation Sequencing Technologies.” *PLoS ONE* 6 (3). doi:10.1371/journal.pone.0017915.

Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael a Gore, Peter J Bradbury, et al. 2010. “Mixed Linear Model Approach Adapted for Genome-Wide Association Studies.” *Nature Genetics* 42 (4): 355–60. doi:10.1038/ng.546.

Zhou, Xiang, and Matthew Stephens. 2012. “Genome-Wide Efficient Mixed-Model Analysis for Association Studies.” *Nature Genetics* 44 (7): 821–24. doi:10.1038/ng.2310.

Articoli pubblicati, sottomessi o in preparazione

Articoli pubblicati

1. Bellucci E, … **Panziera A**, et al. 2014. Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean. Plant Cell 26:1901-12.

2. Benazzo A, **Panziera A**, Bertorelle G. 2015. 4P: Fast computing of population genetics statistics from large DNA polymorphism panels. Ecology and Evolution, 5(1): 172-175

Articoli sottomessi

1. Calderoni L, … **Panziera A**, et al. 2016. Relaxed selective constraints drove functional modifications in the peripheral photoreception of the cavefish P. andruzzii and provides insight into the time of cave colonization. (Submitted to Heredity)

Articoli in preparazione

1. Zanetti E, **Panziera A**, et al. Evidence of strong demographic reduction in European hake populations (Merluccius merluccius)

2. Cornetti L, … **Panziera A**, et al. The transition from oviparity to viviparity: a genomic perspective in the lizard Zootoca vivipara

3. Fuselli S, **Panziera A**, et al. The study of MHC variation and evolution in the Alpine chamois using portable nanopore technology (MinION)

4. Benazzo A, … **Panziera A**, et al. The extreme genomic effects of fragmentation in the Italian brown bear.

Contributi a congressi

1. Cornetti L, … **Panziera A**, et al. NGS approach for investigating evolutionary transition from oviparity to viviparity in squamate reptiles. V Congresso SIBE, Trento 27-30 Agosto 2013.

2. Fuselli S, ... **Panziera A**, et al. Regressive evolution in Somalian Cavefish *Phreatichthys andruzzii*: loss of selective constraint on opsin genes. Congresso SMBE, Vienna 12-16 Luglio 2015.

3. Cornetti L, … **Panziera A**, et al. Genomic insights into the transition from oviparity to viviparity: The case of the reproductively bimodal lizard *Zootoca vivipara*. Congresso ESEB, Lisbona 10-14 Agosto 2015.

3. Benazzo A, … **Panziera A**, et al. The evolution of the small and isolated population of Apennine brown bears (*Ursus arctos marsicanus*): a whole genomes perspective. VI Congresso SIBE, Bologna 31 Agosto – 3 Settembre 2015.

4. Fuselli S, **Panziera A**, et al. The study of MHC variation and evolution in the Alpine chamois: from targeted Sanger sequencing to portable nanopore technology (MinION). VI Congresso SIBE, Bologna 31 Agosto – 3 Settembre 2015.

5. Calderoni L, ... **Panziera A**, et al. Regressive evolution in Somalian Cavefish *Phreatichthys andruzzii*: loss of selective constraint on opsin genes. VI Congresso SIBE, Bologna 31 Agosto – 3 Settembre 2015.

6. Plantard O, … **Panziera A**, et al. Using population genetics to assess tick dispersal, from the mainland to the landscape scale: a review of current knowledge and its utility to design tick-control methods. Genes, Ecosystems and Risk of Infection, Heraklion 21-23 Aprile 2015

Finanziamento del progetto di dottorato

La mia borsa di dottorato è stata co-finanziata dalla Fondazione Edmund Mach di San Michele all'Adige (Trento) e dall'Università degli Studi di Ferrara.