UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DI VITERBO

Dipartimento di Ecologia e
Sviluppo Economico Sostenibile

Dipartimento di
Produzioni Animali

Corso di Dottorato di Ricerca in

ECOLOGIA E GESTIONE DELLE RISORSE BIOLOGICHE
XVIII CICLO

Metodi classici e modelli a coalescente nella ricerca di "selection signatures" tramite marcatori neutrali

Settore scientifico–disciplinare: AGR/17

Coordinatore:
Prof. Giuseppe Nascetti

……..

Tutor:
Prof. Alessio Valentini

……..

Dottorando:
Dott. Stefano Scardigli

……..

Indice

Data consist of number, of course. But these numbers are fed into the computer, not produced by it. These are numbers to be treated with considerable respect, neither to be tampered with, nor subjected to a numerical process whose character you do not completely understand. You are well advised to acquire a reverence for data that is rather different from the "sporty" attitude that is sometimes allowable, or even commendable, in other numerical tasks.

Press, Teukolsky, Vetterling, Flannery:
"Statistical Description of Data"
in NUMERICAL RECIPES p.603
Cambridge University Press, 1992

1

Introduzione

Viviamo in un momento storico nel quale la diversità biologica mondiale viene rapidamente distrutta. Ci sono più specie sulla terra nel tempo geologico attuale di quante ce ne siano mai state in passato ma, come risultato dell'attività umana, anche il ritmo di estinzione delle specie è più grande di quanto lo sia stato in passato.

La perdita di diversità biologica avviene a tutti i livelli: ecosistemi e comunità sono degradati e distrutti e molte specie sono condotte ad estinzione. Questo accade sia nelle aree tropicali che nelle zone temperate, sia negli habitat terrestri che in quelli acquatici. Anche nelle specie che comunque sopravvivono si ha una progressiva perdita di diversità genetica da un lato per la riduzione del numero complessivo di individui che le costituiscono, dall'altro per il crescente isolamento delle popolazioni le une dalle altre e la formazione di isole di consanguineità.

La diversità genetica si sta perdendo anche nelle specie domestiche destinate alla produzione alimentare o comunque industriale dal momento che la produzione abbandona le tecniche di allevamento o coltivazione tradizionale in favore di metodi che tendono a replicare massivamente gli individui che corrispondono a particolari criteri commerciali su larga scala.

1.1 La perdita di biodiversità nelle razze allevate

Il maggiore serbatoio di diversità biologica nelle specie domestiche si trova nella divisione storica in razze, spesso con caratteri estremamente diversificati.

L'uomo ha allevato ad esempio bovini domestici per circa 8500 anni, almeno in Asia minore e nel Medio Oriente. Possiamo partire dall'assunzione che la domesticazione dei bovini sia avvenuta indipendentemente in molti luoghi differenti. Questo è molto interessante dal punto di vista degli allevatori, poiché la domesticazione in diversi luoghi ha così condotto ad una collezione di varie popolazioni profondamente differenti nel loro patrimonio genetico.

Dal momento che solo un piccolo numero di individui selvatici è stato coinvolto in questo processo, la differente provenienza ha comunque contribuito alla ricchezza del patrimonio genetico complessivo delle popolazioni domestiche originarie. Successivi incroci hanno portato a combinazioni non presenti nel tipo selvatico e alla formazione delle razze domestiche attuali.

Nel caso degli animali selvatici la perdita di diversità biologica è essenzialmente dovuta alla scomparsa di ecosistemi e comunque a fattori che riducono progressivamente la consistenza ed il numero di popolazioni di una data specie (a parte la nascita di nuovi ecosistemi come le aree urbanizzate). Nel caso degli animali d'allevamento viceversa non si ha, in generale, una riduzione della numerosità delle popolazioni della specie ma un appiattimento della diversità biologica all'interno della stessa specie. Per le popolazioni allevate si assiste ad un adattamento delle specie al nuovo ecosistema "stalla industriale" e il fitness del fenotipo è "premiato" dalla grande distribuzione commerciale. Il risultato è quindi quello di popolazioni di grandi dimensioni ma estremamente specializzate ed adattate e che, in modo analogo alle specie selvatiche, corrono seri pericoli di estinzione in caso di variazioni del loro habitat anche di lieve entità.

Ad aggravare la situazione delle specie domestiche è l'utilizzo di tecnologie di riproduzione assistita in cui, ad esempio, bovini maschi possono generare nella loro vita riproduttiva decine di migliaia di vitelli.

Inoltre dal momento che l'esportazione è resa più facile dallo sviluppo di nuove biotecnologie, come la crioconservazione del seme e il congelamento degli embrioni, queste razze dominano internazionalmente assumendo una distribuzione geografica sopra-regionale.

La conseguenza in ciò, oltre la possibilità di ampia e rapida diffusione geografica di monotipi genetici ad alta prestazione produttiva, è il rischio di consistente *inbreeding* che

implica ad esempio l'esposizione di intere popolazioni a patologie sia di carattere epidemico sia legate alla comparsa di caratteri recessivi.

Questo, insieme alla condizione di allevamento industriale dove la prossimità forzata degli individui è spinta al massimo e con tecniche come l'utilizzo di farine alimentari di origine animale ad alto contenuto proteico, produce i risultati che spesso compaiono in letteratura anche non strettamente scientifica.

Il numero di razze locali è diminuito costantemente contemporaneamente all'espansione delle razze ad alta produzione. Si può supporre che attualmente la popolazione mondiale di bovini (ca. 250 milioni di capi) sia costituita essenzialmente di solo circa venti razze. Molte delle attuali 500 razze esistenti corrono il pericolo di completa estinzione o di perdita dei loro caratteri peculiari per l'incrocio con razze dominanti [1].

Le ragioni per la conservazione della biodiversità, nel caso di animali d'allevamento costituita dalla specificità delle razze, sono molteplici.

Le razze in pericolo di estinzione possono contenere tratti genetici non osservati o sconosciuti. Questi possono essere vantaggiosi rispetto ai tratti esistenti e predominanti nelle popolazioni attuali in previsione di cambiamenti ambientali, variazioni dei mercati, o programmi di *cross-breeding* con altre popolazioni. Tali vantaggi derivano sia dalla presenza di singoli geni che andrebbero inevitabilmente persi con la sparizione di una particolare razza, sia essere il risultato dell'interazione di diversi geni e sarebbe molto difficile e costosa la ricostruzione di quella particolare combinazione a partire dal patrimonio di geni delle razze a maggiore diffusione.

Ancora, è impossibile prevedere la domanda futura dei mercati e i cambiamenti nei sistemi di produzione. E' quindi da supporre che i caratteri degli animali possano differire in modo considerevole in futuro da quelli attuali.

In condizioni ambientali estreme, come ad esempio quelle che danno luogo a nomadismo, razze opportunamente adattate con una bassa produttività ma anche con una minimale necessità per la sussistenza, possono offrire molti vantaggi rispetto a razze esotiche che richiedono complesse tecniche di allevamento o di alimentazione a causa delle loro prestazioni produttive.

Da non tralasciare infine l'aspetto culturale della diversificazione delle razze in quanto intimamente legato alla storia umana e in grado comunque di costituire una risorsa

economica se opportunamente gestito per lo sfruttamento ad esempio nel mercato di prodotti ad alta tipicità o in campo turistico.

E' quindi estremamente importante lo studio e il controllo dei programmi di selezione delle specie d'allevamento al fine di controllare che le esigenze economiche non portino all'estremo i processi di selezione fino a mettere in condizione di rischio le specie in questione e gli stessi consumatori.

D'altra parte lo studio delle specie d'allevamento, come ad esempio i bovini, risulta estremamente interessante come laboratorio dove sperimentare tecniche di analisi statistica da applicare alle popolazioni di specie selvatiche.

Infatti le mutazioni degli ecosistemi artificiali costituiti dagli allevamenti, ovvero le mutazioni dei mercati e delle esigenze economiche, sono generalmente molto più rapide di quelle che avvengono negli ecosistemi naturali. Il confinamento, deprecabile per altri versi, costituisce uno strumento di sicura identificazione delle popolazioni e degli individui. E' spesso ben documentata la storia della discendenza degli individui, della costituzione di popolazioni e della formazione di razze. Non ultima, la disponibilità di campioni di individui è sicuramente molto superiore sia come numero, sia come tipologia di quella degli individui selvatici.

1.2 L'obiettivo del presente lavoro

Lo scopo del presente lavoro è quello di mettere a confronto metodi per la ricerca di siti di selezione genetica basati sulla statistica di popolazione utilizzando come caso di studio campioni da popolazioni allevate artificialmente. Per questo si è scelto un gene particolarmente sfruttato nelle popolazioni bovine e deputato alla sintesi della miostatina. Questa proteina interviene come moderatore nello sviluppo muscolare durante la gestazione e la crescita dei vitelli. Un difetto genetico che porti alla mancata o ridotta produzione della miostatina porta al fenotipo cosiddetto "doppia coscia", caratterizzato da un anormale incremento delle masse muscolari del soggetto interessato da questa disfunzione. Chiaramente l'interesse commerciale di individui con un difetto in questo gene è alto e alcune razze in cui diverse mutazioni alterano la funzionalità di questo gene hanno subito in passato e subiscono attualmente una notevole pressione selettiva.

I metodi di indagine impiegati saranno prima verificati su una popolazione simulata che riproduce le caratteristiche essenziali di una popolazione analoga a quelle sotto selezione per il gene della miostatina. Successivamente gli stessi metodi verranno applicati al caso di studio reale e sarà possibile confrontare i risultati con quelli ottenuti nella simulazione.

La storia della mutazione di questo gene è registrata in letteratura e negli herd-book delle razze in questione. Non si vuole quindi, con questo lavoro, aggiungere nuove conclusioni sulla diffusione della mutazione del gene della miostatina, quanto piuttosto evidenziare le possibilità e i limiti che metodi di indagine statistica presentano. I campioni di misura considerati presentano anche un ulteriore grado di problematicità in quanto sono piuttosto limitati in numero e nel caso di una delle razze in questione non sono rappresentativi della distribuzione della mutazione nella popolazione.

Si vedrà come metodi statistici completamente differenti possano cooperare a formare un quadro unitario del problema in esame e, nonostante la situazione decisamente sfavorevole, sia comunque possibile evidenziare gli effetti di selezione su questa mutazione giungendo a conclusioni che sono perfettamente in linea con le fonti storiche.

2

Ricerca di “selection signatures” tramite marcatori neutrali

Una delle possibili strategie per scoprire nelle sequenze di DNA siti soggetti a selezione è l'analisi di marcatori neutrali provenienti dalle regioni di DNA dove si sospetta la presenza del sito. Se si osserva una distribuzione statistica di un qualche marcatore che differisce da quelle degli altri osservati, allora si può supporre che tale marcatore sia in prossimità dell’area soggetta a selezione. Infatti, ci si può aspettare che le dinamiche di popolazione e le strutture demografiche facciano sentire il loro effetto allo stesso modo su tutti i marcatori, mentre la presenza di un sito di selezione induca un comportamento differente solo in alcuni dei marcatori [2].

Diverse tecniche statistiche che fanno uso delle frequenze alleliche e dell’eterozigosità di singoli loci sono utilizzate negli studi di selezione. Tra queste saranno prese in considerazione l’analisi di Linkage Disequilibrium e l’analisi dell’indice di fissazione Fst. Tali metodi saranno poi confrontati con i risultati di un approccio completamente differente derivato da una formulazione elementare della teoria del coalescente.

Questo capitolo presenta il substrato teorico dei metodi statistici utilizzati nel presente lavoro.

2.1 Il Linkage Disequilibrium

Se si considera una coppia di loci biallelici (*A*-*a*, *B*-*b*) in una popolazione e gli alleli sono casualmente distribuiti (come nel caso ci si aspetta di marcatori neutrali), la probabilità di avere un alplotipo (*Ab*) è semplicemente il prodotto della probabilità di avere un allele *A* nel primo locus moltiplicata la probabilità di avere un allele (*b*) nel secondo

$$P(Ab)=P(A)P(b)$$

Al fine di misurare lo stato di equilibrio statistico delle distribuzioni di alleli si può introdurre un indicatore **D** che misura la differenza tra la probabilità composta e il prodotto delle probabilità indipendenti:

$$\mathbf{D} = \mathrm{P}(Ab) - \mathrm{P}(A)\mathrm{P}(b)$$

Se l'aplotipo Ab si trova sullo stesso frammento di DNA (ovvero è in linkage) di una mutazione soggetta a selezione positiva, si osserva nella popolazione un eccesso di tali aplotipi e il valore di **D** risulta maggiore di zero e si parla di linkage disequilibrium (LD) tra i marcatori.

Inoltre, l'ammontare di LD è inversamente correlato alla distanza dei marcatori, dal momento che selezione ed eventi di ricombinazione agiscono in competizione e la probabilità di un evento di ricombinazione è approssimativamente proporzionale alla distanza sul frammento dei marcatori stessi.

La distanza di marcatori sulla sequenza di DNA viene infatti espressa anche nell'unità "centiMorgan" (cM): 1 cM corrisponde alla distanza tra due basi per le quali la probabilità di ricombinazione in una meiosi è pari all'uno percento. Mediamente il cM corrisponde ad una distanza di 1 milione di basi. Da notare che due basi che si trovino in due cromosomi differenti hanno una probabilità di "ricombinazione", ovvero di essere scambiati tra due cromosomi omologhi, pari al 50 percento. Chiaramente è privo di senso dire che si trovano a 50 cM di distanza ma questo esempio evidenzia come la misura di distanza in cM abbia essenzialmente un valore locale.

Sulla scala temporale dei processi di selezione artificiale, come nel caso reale studiato in questo lavoro, è possibile trascurare eventi di mutazione. I processi di ricombinazione sono quindi l'unico motore in grado di ristabilire l'equilibrio statistico degli alleli.

L'indice di LD che sarà usato in questo lavoro è il **D'** di Lewontin normalizzato [3][4]. Se si considerano due loci multiallelici e si indicano con p_i e q_j le frequenze degli alleli i e j dei rispettivi loci, allora tale indice può essere espresso come

$$D' = \Sigma \Sigma \, p_i \, q_j \, |D_{ij} / D_{max}|$$

dove le somme sono estese su tutti gli alleli del primo e del secondo locus, e dove

$$D_{max} = \min [\, p_i q_j, (1-p_i)(1-q_j)] \qquad \text{se} \qquad D_{ij}<0$$

$$D_{max} = \min [\, p_i(1-q_j), (1-p_i)q_j] \qquad \text{se} \qquad D_{ij}>0$$

D_{max} rappresenta il coefficiente di normalizzazione legato al valore delle frequenze dei singoli alleli e garantisce l'indipendenza dall'indice dalle frequenze alleliche.

2.2 L'indice di fissazione Fst

Gli indici di fissazione Fis, Fit e Fst, introdotti da Wright nei primi anni '50, offrono un modo conveniente di rappresentare la struttura di una popolazione e la sua suddivisione in sottopopolazioni. I tre indici possono essere visti come la suddivisione in tre componenti della varianza allelica di un dato gene nell'intera popolazione. Mentre Fis e Fit rappresentano la correlazione tra due alleli prelevati da una popolazione in relazione rispettivamente ad una sottopopolazione e all'intera popolazione, Fst rappresenta la probabilità di estrarre casualmente da una sottopopolazione due alleli diversi rispetto alla stessa probabilità sull'intera popolazione.

I tre indici sono collegati dall'espressione [5]

$$Fis = (Fit - Fst) / (1 - Fst)$$

Qualora la popolazione totale non abbia una reale suddivisione in sottopopolazioni chiaramente

$$Fst = 0 \quad => \quad Fis = Fit$$

La stima classica di Wright dell'indice di fissazione Fst, e che comprende il principio di Wahlund, è quella che si ottiene come

$$Fst = \sigma_o^2 / p_t(1-p_t) \qquad (A)$$

dove σ_p^2 è la varianza della frequenza allelica nelle sottopopolazioni e $p_t(1-p_t)$ è proporzionale all'eterozigosità aspettata per l'intera popolazione.

Questa stima deriva da un modello ideale di sottopopolazioni discendenti in modo parallelo da un'unica popolazione ancestrale. La maggior parte delle popolazioni reali invece presentano una storia di tipo filogenetico e per questo Nei [6] introduce una diversa stima dell'Fst ottenuta direttamente dall'eterozigosità osservata nelle sottopopolazioni h_o e dall'eterozigosità aspettata h_s nell'intera popolazione:

$$Fst = 1 - h_o/h_s \qquad (B)$$

La ricerca degli effetti di selezione avverrà, come nel caso del LD, dal confronto del comportamento dell'Fst dei diversi microsatelliti nelle diverse popolazioni.

Mentre la stima (B) sarà calcolata direttamente dalle misure dei campioni in esame per la stima (A) si utilizzerà la *suite* di software **fdist2** [7] che utilizza l'estensione corretta per gli effetti di *bias* per il calcolo del rapporto delle varianze.

Inoltre, in questo secondo caso, lo stesso software mette a disposizione un programma di simulazione basato sul coalescente (cfr. paragrafo 2.3) che permette di stimare l'Fst in un modello teorico "a isole" come funzione dell'eterozigosità aspettata. Questo modello considera un set di marcatori neutrali e fornisce una banda di confidenza per l'identificazione di *outliers*. L'esame dei marcatori osservati sarà quindi condotto non per confronto tra il loro diverso comportamento relativo ma in funzione del loro posizionarsi o meno nel intervallo di confidenza stimato dalla simulazione. Gli *outliers* rispetto alla distribuzione teorica potranno essere identificati come marcatori prossimi al sito di selezione.

2.3 Il coalescente

Consideriamo un frammento di DNA e supponiamo sia sufficientemente corto da poter trascurare eventi di ricombinazione al suo interno. Secondo il modello a coalescente [8][9] si può assumere che se si va indietro nel tempo per un numero sufficiente di generazioni tutti gli omologhi frammenti di DNA provenienti dagli individui di un dato campione di una popolazione discendano da un singolo frammento appartenuto ad un progenitore comune, il *most recent common ancestor* o MRCA (fig. 2.3.1).

Lo studio della storia del campione, relativamente al dato frammento di DNA, deve quindi essere riconsiderato come lo studio dell'albero di coalescenza che unisce gli individui del campione al MRCA, attraverso un imprecisato numero di individui intermedi non osservati. Se nei frammenti del campione si riscontrano differenti alleli per alcuni loci si deve concludere che nell'albero di coalescenza siano comparse mutazioni trasmesse nei frammenti appartenenti agli individui discendenti dallo stesso ramo.

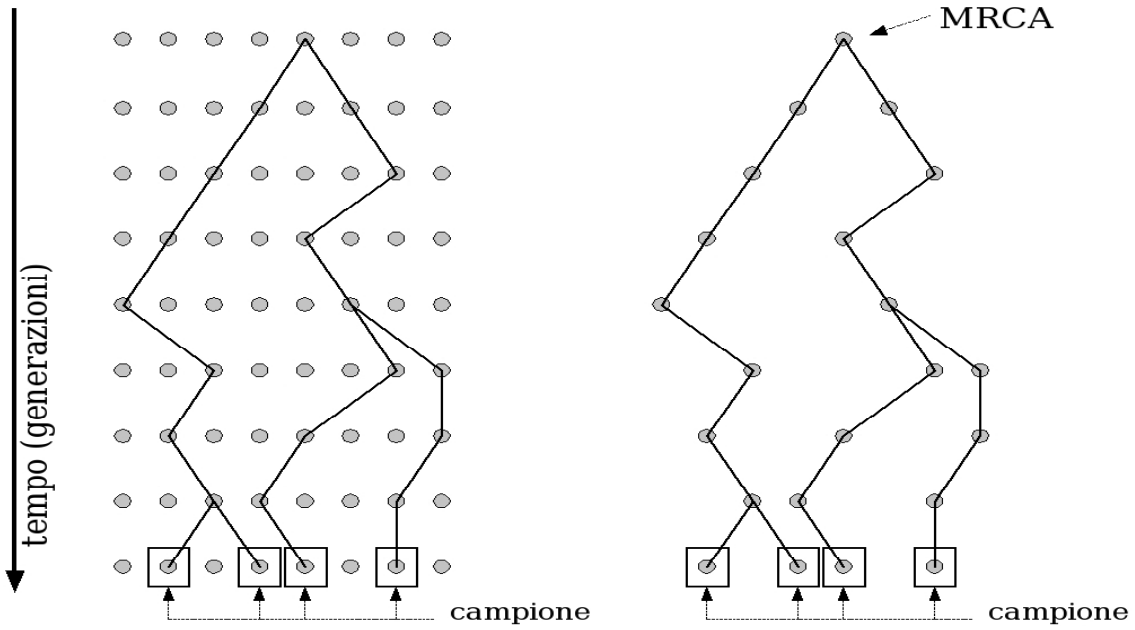


Fig. 2.3.1: albero di coalescenza relativo ad un campione di popolazione

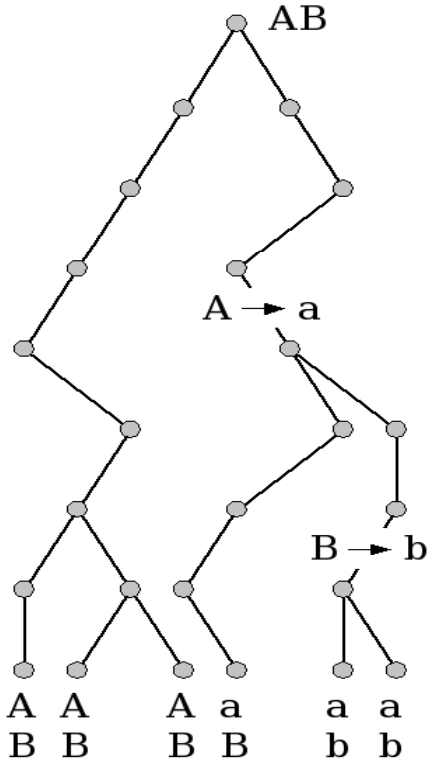


Fig. 2.3.2: albero di coalescenza con la comparsa di mutazioni

Nella figura 2.3.2 si vede come la comparsa di due mutazioni in due loci producano una distribuzione di aplotipi (AB, aB, ab) a partire da un MRCA di aplotipo (AB).

Se, inoltre, nella popolazione si considerano solo marcatori neutrali, se si assume che la dimensione della popolazione non vari nel tempo e che le dinamiche di riproduzione non alterino l'equilibrio Hardy-Weimberg (HWE), allora la storia di coalescenza del campione è ben descritta dal *mutation rate* μ dei marcatori, dalla dimensione effettiva N_e e dalla topologia dell'albero.

Mentre μ e N_e sono parametri dell'intera popolazione, la topologia dipende solo dallo specifico campione in esame. Dal momento che l'interesse è rivolto allo studio della popolazione è possibile dimenticarsi della topologia e concentrarsi solo su tali parametri.

Sfortunatamente, applicando il modello a coalescente, non è possibile stimare μ e N_e indipendentemente ma solo il *rescaled mutation rate* Θ definito come

$$\Theta = \alpha \mu N_e$$

dove α è una costante che tiene conto della ploidità degli individui (esm. $\alpha = 2$ per individui aploidi, $\alpha = 4$ per individui diploidi, …).

L'aggettivo "rescaled" che si dà a questo parametro deriva dalla considerazione seguente. La scala dei tempi più naturale in un modello di coalescenza è il numero di generazioni che possiamo indicare con t. Il numero totale m_{tot} di mutazioni di un dato locus per tutti gli individui del campione e che sono avvenute ripercorrendo all'indietro nel tempo il processo di coalescenza fino al MRCA è dato semplicemente dal prodotto del *mutation rate* μ di tale locus per la somma t_{tot} delle lunghezze temporali di tutti i rami dell'albero:

$$m_{tot} = \mu t_{tot}$$

Se si introduce un "tempo effettivo" τ definito come

$$\tau = t / N_e$$

Allora si può scrivere

$$m_{tot} = \mu t_{tot} = \mu N_e t_{tot} / N_e \sim \Theta \tau_{tot}$$

Quindi si può interpretare Θ come il *mutation rate* per il dato locus riferito al tempo effettivo τ.

Essendo però interessati alla stima separata di mutation rate e popolazione effettiva non *rescaled* si può comunque procedere per confronti. Se si comparano gli stessi marcatori in popolazioni differenti o differenti marcatori nella stessa popolazione, il rapporto dei Θ fornisce una stima dei relativi N_e e dei relativi μ.

Ad esempio consideriamo due popolazioni P e Q e due marcatori A e B, si avranno in totale quattro determinazioni di Θ:

	Locus A	Locus B
Popolazione P	Θ_{PA}	Θ_{PB}
Popolazione Q	Θ_{QA}	Θ_{QB}

Da cui potremmo dedurre i rapporti

$$N_{eP} / N_{eQ} = \Theta_{PA} / \Theta_{QA} = \Theta_{PB} / \Theta_{QB}$$

e

$$\mu_A / \mu_B = \Theta_{PA} / \Theta_{PB} = \Theta_{QA} / \Theta_{QB}$$

Il modello a coalescente cambia anche concettualmente la visione della misura di marcatori in una data popolazione, introducendo la storia del particolare campione prelevato come fattore importante[10]. Nel caso rappresentato nella figura 2.3.2 si hanno le seguenti frequenze aplotipiche:

AB	Ab	aB	ab
0.50	0.17	0.00	0.33

Semplicemente scambiando tra loro gli eventi di mutazione, si ottengono frequenze leggermente diverse:

AB	Ab	aB	ab
0.50	0.00	0.17	0.33

Lo scambio è assolutamente lecito in quanto si considerano gli eventi di mutazione assolutamente casuali e i marcatori completamente neutrali. Questo semplice caso registra un effetto poco appariscente a causa del piccolo numero di individui del campione, ma mostra come, nel caso di un numero di individui più consistente, la particolare storia del campione possa alterare considerevolmente analisi basate, ad esempio, sulle frequenze aplotipiche.

La constatazione che a formare il campione esaminato concorrano solo un numero limitato di individui progenitori fino al MRCA ha indotto anche una innovazione nelle strategie utilizzate nelle simulazioni numeriche di popolazioni.

Nei modelli classici di simulazione del tipo Wright-Fisher un *pool* di individui viene fatto evolvere in generazioni successive secondo varie regole atte a simulare il

comportamento di una popolazione reale. Dallo stato finale o da stati intermedi della simulazione si possono estrarre campioni da studiare in base alle strategie utilizzate.

L'affidabilità della simulazione e la confidenza dei parametri statistici ottenuti, derivano dalla ripetizione della simulazione un numero molto grande di volte, con notevoli costi in termini di tempo di calcolo.

Con l'introduzione del modello a coalescente sono nate tecniche di simulazione basate sulla generazione diretta di alberi di coalescenza e dei relativi campioni incomparabilmente più rapide delle simulazioni classiche. Ne è un esempio il software fdist2, già citato in merito all'indice Fst, che si basa su una simulazione di questo tipo ed è in grado di generare 20000 campioni indipendenti in alcuni minuti su un personal computer.

Per la ricerca di indicatori di selezione su marcatori neutrali, si utilizzerà il software LAMARC [11] che stima i valori di Θ per i campioni di popolazioni tramite una ricerca di maximum likelihood.

Dal momento che è possibile supporre infatti che lo stesso marcatore neutrale abbia lo stesso μ in popolazioni differenti, quando il rapporto dei Θ stimati di campioni di due popolazioni differisce, ovvero

$$\Theta_{PA} / \Theta_{PB} \neq \Theta_{QA} / \Theta_{QB}$$

si potrà concludere che siano in corso differenti dinamiche nelle popolazioni rispetto ai due marcatori.

In altre parole, dal momento che la formulazione elementare del modello a coalescenza descrive la dinamica di marcatori neutrali in popolazioni che ben si comportano dal punto di vista statistico (*random-mating*, dimensione quasi infinita, etc.) è possibile utilizzare il coalescente come un modello 0 e assumere anomalie dei Θ stimati come deviazioni dall'ipotesi nulla di nessuna selezione.

Quello che ci si aspetta è che un dinamica di selezione positiva operi nel ridurre la ricchezza del *pool* di alleli di una popolazione. In un modello di coalescenza un numero ridotto di alleli risulta in un albero di coalescenza con un numero stimato di generazioni più piccolo dal campione osservato al MRCA. Viceversa marcatori a maggiore distanza dal sito di selezione saranno rappresentati da alberi di coalescenza più lunghi.

2.4 Ricostruzione degli aplotipi

Nel caso di soggetti diploidi, generalmente si indica con il termine "fenotipo" l'insieme degli alleli che il pool di marcatori presenta in un dato individuo, senza riferimento alla posizione relativa sui due cromosomi omologhi, con il termine "genotipo" invece lo stesso insieme di alleli ma identificando la posizione relativa, o "fase", dei diversi alleli sui due cromosomi omologhi. In questo secondo caso si parla di "aplotipo" per identificare l'insieme di alleli provenienti da un unico cromosoma.

Nello studio del linkage disequilibrium non è sufficiente la conoscenza delle frequenze alleliche del pool di marcatori considerati ma è necessario conoscere gli esatti aplotipi proprio perché si è alla ricerca della storia comune di alleli che hanno "viaggiato insieme".

Purtroppo ogni qualvolta l'osservazione degli alleli di marcatori viene condotta con la procedura standard PCR+gel elettroforetico, questo tipo di analisi perde l'informazione sull'appartenenza di un dato allele ad uno dei due cromosomi omologhi.

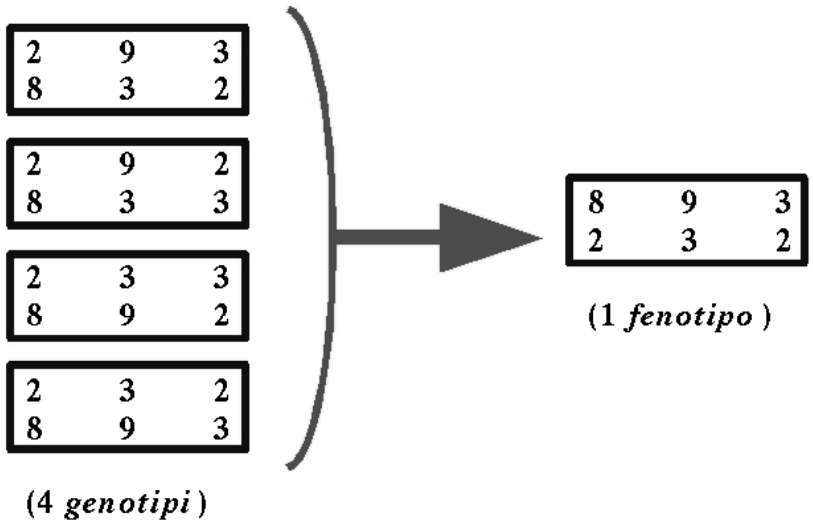


Fig. 2.4.1 : ipotetica analisi PCR+gel di quattro genotipi

Nella fig. 2.4.1 è rappresentato un esempio ipotetico del caso di quattro individui ciascuno identificato da tre marcatori multiallelici. Il primo individuo ha due frammenti di DNA contenenti rispettivamente gli alleli (2,9,3) e (8,3,2). Tutti è quattro gli individui hanno una diversa disposizione di alleli e quindi hanno 4 genotipi differenti e il *pool* di marcatori presenta un totale di 8 aplotipi distinti. L'analisi tramite PCR+gel produce la

separazione degli alleli riferiti a ciascun marcatore e il loro ordinamento per peso molecolare decrescente.

Quindi per tutti gli individui il risultato dell'analisi è identico ovvero presentano tutti lo stesso fenotipo e si perde qualsiasi informazione sugli aplotipi del campione.

Una possibilità di recuperare questo dato si ha quando sono disponibili informazioni parentali degli individui e procedendo quindi per confronti, individuo per individuo. Nel caso del presente lavoro non sarà disponibile questa informazione ed è stato quindi necessario ricorrere alla ricostruzione degli aplotipi su base statistica.

Esiste una vasta letteratura [12][13][14] sulle tecniche che rendono possibile la ricostruzione di una distribuzione presunta degli aplotipi del campione a partire dai fenotipi osservati sperimentalmente. Nel presente lavoro si farà uso del software PHASE 2.1.1 [15] e la fig. 2.4.2 rappresenta un esempio di come il metodo operi su un campione ipotetico di 9 individui diploidi.

Nella tabella di sinistra sono riportati gli alleli "misurati" e "reali" del campione: nella colonna di sinistra vi sono i diciotto aplotipi "reali", nella colonna centrale vi sono gli alleli

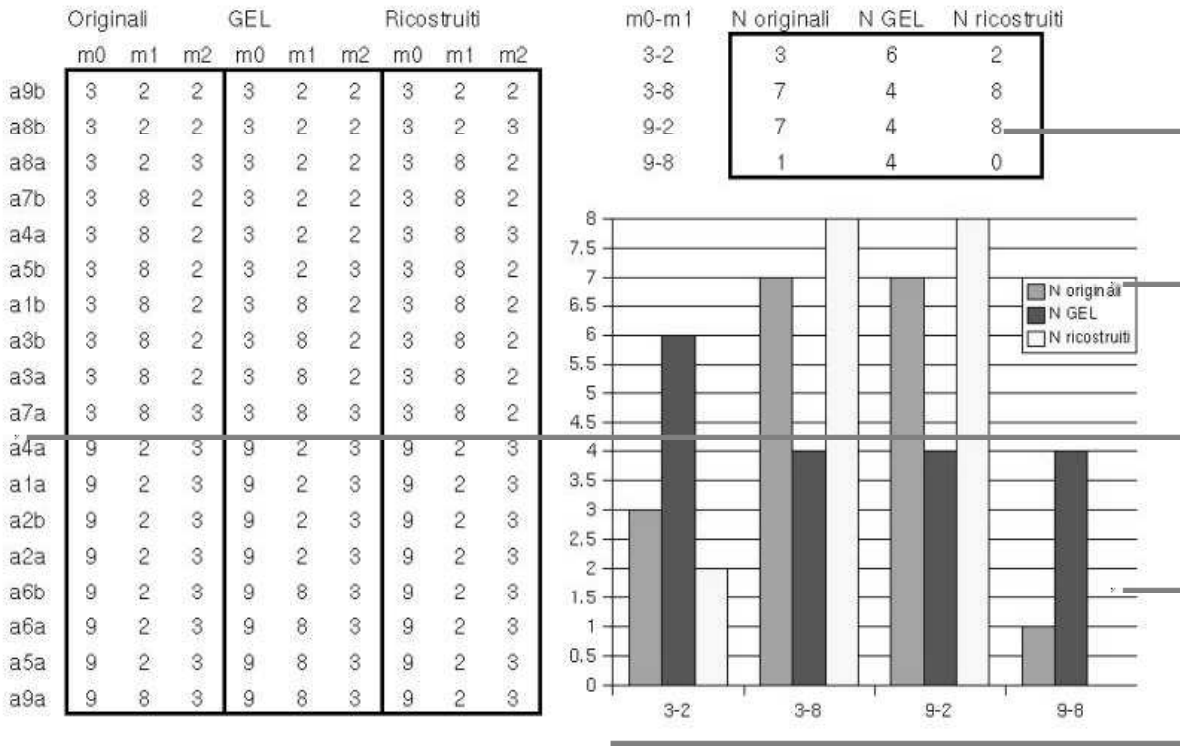


Fig. 2.4.2 : ricostruzione degli aplotipi con il software PHASE 2.1.1

osservati "sperimentalmente" e attribuiti ai diversi individui e nella colonna di destra gli aplotipi ricostruiti dal software.

Nell'istogramma a destra, nella stessa figura, è rappresentato il numero di aplotipi differenti per i primi due marcatori (m0 e m1) corrispondenti agli aplotipi (3-2), (3-8), (9-2) e (9-8). Il campione originale è costituito da 3 frammenti di DNA del tipo (3-2), 7 frammenti del tipo (3-8), 7 frammenti del tipo (9-2) e 1 solo frammento del tipo (9-8).

La misura "sperimentale" dei marcatori, qualora si usasse direttamente come espressione degli aplotipi, indurrebbe una distribuzione completamente errata: 6 frammenti di DNA del tipo (3-2), e 4 frammenti ciascuno dei tipi (3-8), (9-2) e (9-8). In particolare la presenza dell'ultimo aplotipo risulterebbe ampiamente sovrastimata.

Nello stesso grafico si vede invece come gli aplotipi ricostruiti dal software mostrino una distribuzione non perfetta ma abbastanza vicina a quella "reale".

3

Simulazione di una popolazione sotto selezione

3.1 Il software *Buttero*

Questo software è nato con lo scopo di simulare il comportamento di un pool di marcatori microsatelliti in una o più popolazioni d'individui in un modello Wright-Fisher che includesse però anche la possibilità di estrarre campioni dalle popolazioni e soprattutto di combinarle al fine di studiare situazioni di *admixture* o di isolamento.

Nel corso dello sviluppo del presente lavoro, *Buttero* si è evoluto dal progetto iniziale per includere anche la possibilità di simulare condizioni di selezione e dinamiche di ricombinazione. Per questo motivo, oltre ai marcatori neutrali per i quali è stata introdotta la possibilità di ricombinazione, è stato aggiunto al "patrimonio genetico" degli individui anche un marcatore non neutrale (biallelico) che segnasse la presenza o meno di una mutazione sulla quale far agire la selezione.

Il linguaggio scelto per la realizzazione di *Buttero* è Java e consiste in una serie di classi che possono essere eseguite sia come applet sfruttando la macchina virtuale di un browser HTML, sia come applicativo singolo in una macchina virtuale Java (JVM).

La scelta di questo linguaggio di programmazione è dovuta essenzialmente a due fattori: alla indipendenza dal sistema operativo di questo linguaggio e alla filosofia "ad oggetti" dello stesso. In particolare quest'ultima caratteristica ha permesso lo sviluppo di codice semplice e flessibile permettendo di concentrare l'attenzione sulle singole classi e sui relativi metodi secondo gli standard della Programmazione Orientata agli Oggetti (*Object-Oriented Programming, OOP)*. La felice scelta iniziale si è rivelata utilissima nel progressivo ampliamento delle possibilità del modello iniziale relativamente semplice.

La classi principali di questo modello sono: DNA, Mucca e Branco. Chiaramente i nomi risentono dell'argomento specifico della ricerca in corso, ma nulla è tolto alla genericità del modello.

- **la classe DNA**

Questa classe è ovviamente il modello di sequenza di DNA che caratterizza geneticamente gli individui. Essa è costituita, allo stato attuale, da dieci marcatori microsatelliti per ciascuno dei quali è configurabile un massimo e minimo di lunghezza e il *mutation rate*. Al fine di studiare gli effetti di ricombinazione i microsatelliti sono ordinati (numerati da 0 a 9) e due variabili controllano lo stato di linkage e il coefficiente di ricombinazione con il microsatellite precedente. Chiaramente per il microsatellite 0 queste variabili non vengono utilizzate nella simulazione.

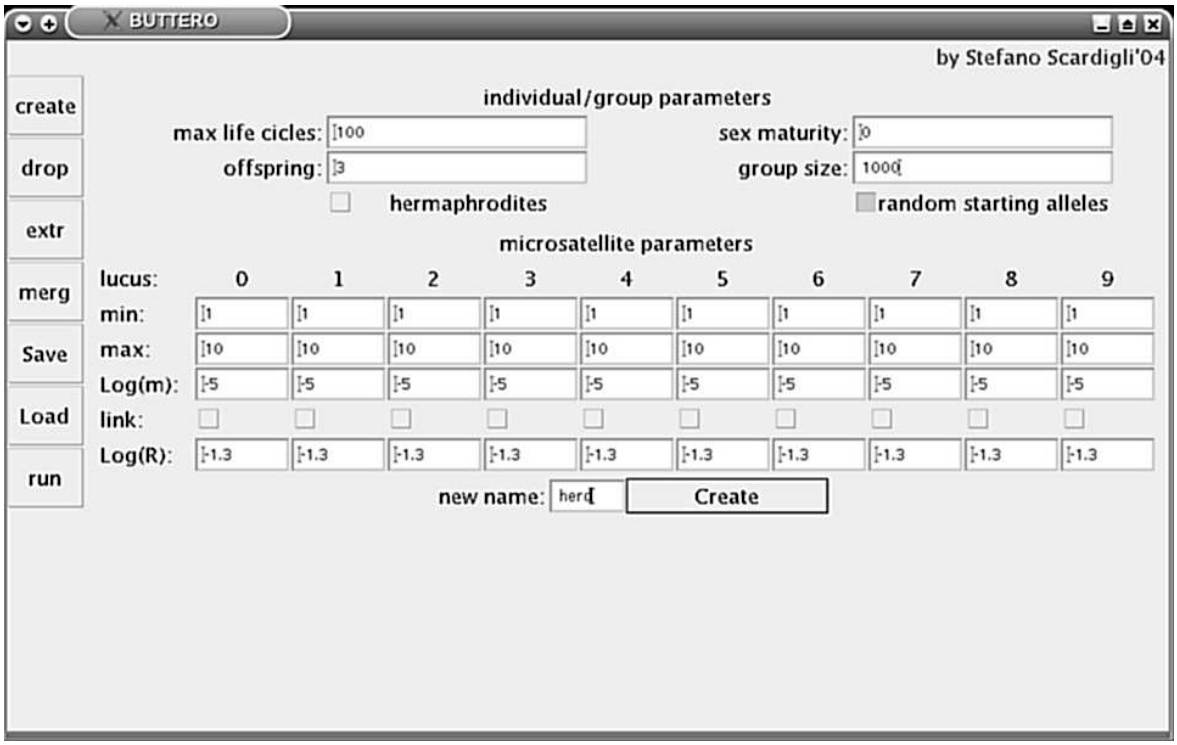


Fig. 3.1.1: pannello di impostazione per il DNA e la popolazione

Il locus soggetto a mutazione è considerato completamente in linkage con il microsatellite 0. Per questo locus non esiste *mutation rate* in quanto durante la simulazione è l'operatore a decidere il momento in cui far comparire la mutazione mentre il modello sceglie casualmente il singolo individuo che ne sarà oggetto.

I metodi di questa classe sono semplicemente quelli necessari ad impostare lo stato dei marcatori, come la lunghezza dei microsatelliti e lo stato di mutazione o meno del locus.

- **la classe Mucca**

Più interessante è la classe che rappresenta l'individuo nel modello. Questa classe oltre a contenere al suo interno due istanze della classe DNA (due cromosomi omologhi), contiene sia attributi statici quali il sesso e l'eventuale ermafroditismo, l'età di maturazione sessuale, l'età massima raggiungibile dall'individuo, la numerosità della prole, sia attributi dinamici quali l'età e la condizione di gravidanza.

- **la classe Branco**

Questa classe costituisce l'unità operativa di *Buttero*, è costituita da una collezione di istanze della classe Mucca e dispone dei metodi necessari all'evoluzione della popolazione come Branco.Popola() che genera una popolazione secondo parametri assegnati, Branco.Tempo() che esegue un ciclo riproduttivo per la popolazione, Branco.GENmuta() che induce la mutazione in uno degli individui della popolazione.

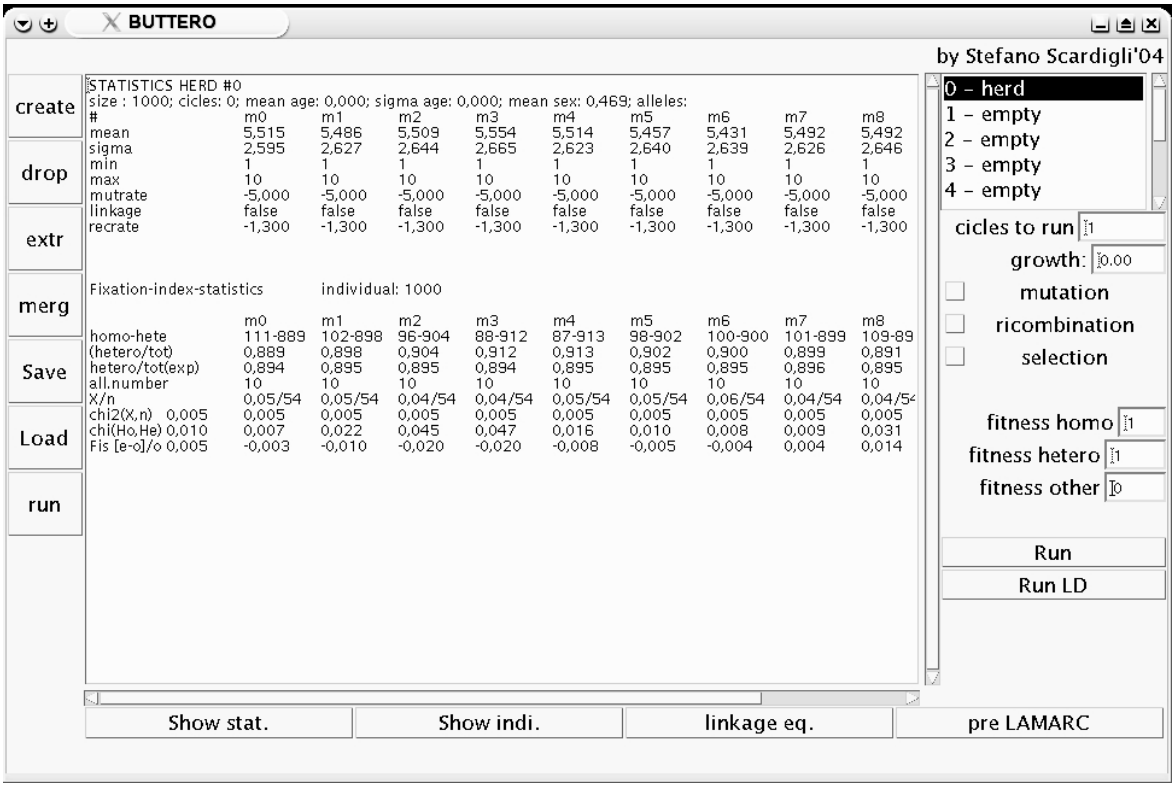


Fig. 3.1.2: pannello di controllo delle simulazioni

Oltre a questi vi sono metodi che restituiscono informazioni statistiche elementari sulla popolazione (età media, rapporto maschi/femmine, eterozigosità, indici di linkage disequilibrium, fitness medio, etc.).

Lo sviluppo di *Buttero* segue in modo parallelo il lavoro di ricerca e quindi non ha ancora un assetto definitivo. Tra gli indirizzi di sviluppo futuri vanno segnalati:

- linguaggio di controllo *batch* del software: l'impostazione originale è quella di un software gestibile in modo interattivo tramite un'interfaccia grafica. E' però in fase di ultimazione lo sviluppo di una classe in grado di interpretare comandi inviati tramite un file *batch* e di eseguirli in successione, non richiedendo l'intervento di operatore. Questo è molto utile nel caso di simulazioni molto lunghe o ripetitive.
- simulazione di individui su "lattice": si tratta di introdurre due coordinate spaziali per gli individui di una popolazione al fine di studiare la diffusione e le fluttuazioni spaziali di una mutazione nella popolazione.

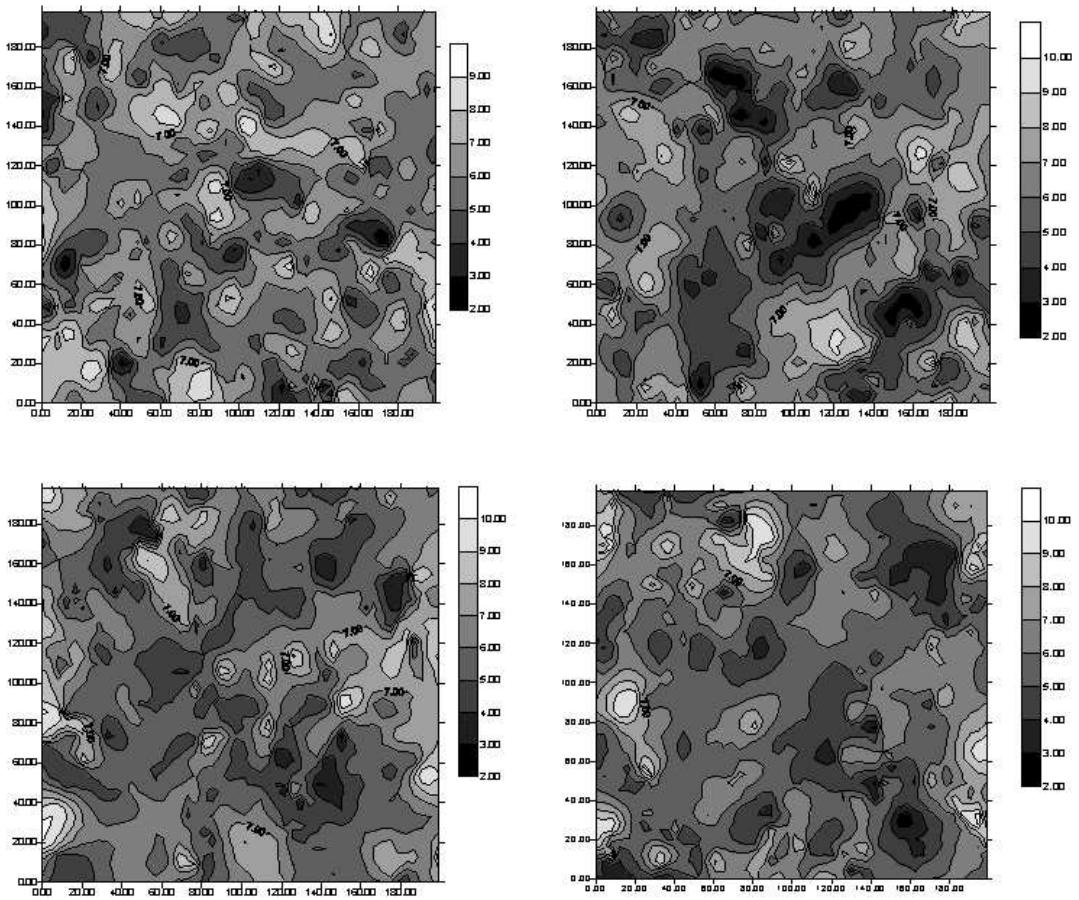


Fig. 3.1.3: quattro distribuzioni spaziali di alleli di microsatelliti

La fig. 3.1.3 è un primo esempio di distribuzioni spaziali di alleli ottenute tramite simulazione su lattice condotta con le nuove versioni preliminari di *Buttero*. Questa parte è ancora in una fase iniziale in quanto implica la riscrittura di diversi metodi e la parziale modifica di alcune classi fondamentali.

3.2 Simulazione di una popolazione sotto selezione

La popolazione che si è scelta per la simulazione consiste di 4000 individui gonocorici. I dieci microsatelliti e il locus di selezione sono stati disposti in modo da trovarsi ad una distanza di 2 cM dal successivo, come rappresentato nella tabella seguente

	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9
Distanza dal locus di selezione (cM)	**0**	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**

Uno dei problemi fondamentali per questo tipo di simulazioni è la scelta delle condizioni iniziali per gli alleli dei microsatelliti. Una distribuzione iniziale completamente casuale di alleli (*hot start*), oltre che non realistica, induce un lungo processo di assestamento con l'induzione di un forte linkage disequilibrium. L'interpretazione del fenomeno risulta piuttosto difficile perché diverse possono essere le cause, come ad esempio effetti "di bordo" dovuti al limite imposto alla lunghezza massima e minima del microsatellite che si traduce in un effetto di selezione sugli alleli estremi. Non si può escludere anche una risposta del linkage disequilibrium analoga a quella che si ha in condizioni di *admixture*. Essendo inoltre una simulazione su una popolazione limitata vi è una naturale tendenza nel modello alla fissazione degli alleli e questo impedisce un lungo (numero generazioni > 200) *warm-up* a partire da tali condizioni iniziali.

Si è preferito quindi imporre una condizione iniziale di alleli completamente fissati (*cold start*) e un periodo di *warm-up* di 30 generazioni con elevato *mutation-rate* (μ=0.1) per simulare una storia genetica abbastanza "realistica" della popolazione, anche se concentrata nel tempo. In fig. 3.2.1 è riportato il grafico delle frequenze alleliche al termine di un *warm-up* di questo tipo.

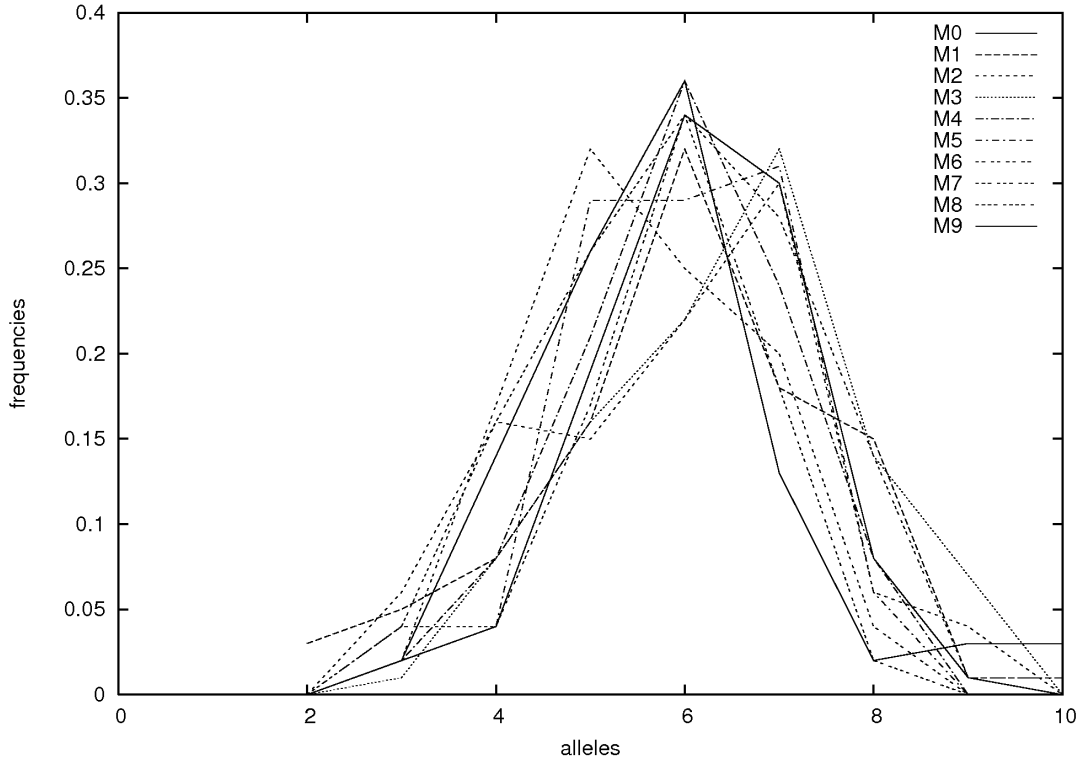


Fig. 3.2.1 frequenze alleliche dopo il *warm-up*.

Nel presente lavoro si è scelto di simulare un carattere dominante piuttosto che uno recessivo, come avrebbe dovuto essere per il caso del gene mutato della miostatina. Ciò deriva dalla limitatezza implicita nella simulazione numerica di una popolazione di dimensione limitata. La selezione su di un carattere recessivo implica una rapidissima affermazione della mutazione nella popolazione simulata, fino alla sua fissazione. Tale rapidità implica la fissazione di praticamente tutti i microsatelliti vicini alla mutazione.

È possibile rappresentare in individui diploidi l'idoneità fenotipica ad una pressione selettiva, o fitness, come un vettore di tre elementi indicanti l'idoneità dell'individuo (nell'ordine):

- omozigote, ovvero potatore in entrambe i cromosomi dell'allele sotto selezione
- eterozigote, quando possiede un solo cromosoma con l'allele in questione
- altro (o *wild*), quando l'allele risulta assente

Ad esempio si indicherà con [2;1;0] un carattere codominante dove il numero di alleli presenti si ripercuote in modo quantitativo nell'espressione del fenotipo. Incidentalmente, i valori assoluti assegnati ai diversi genotipi sono legati alla particolare normalizzazione

scelta e lo stesso carattere potrebbe essere indicato anche come [1;0.5;0] o [6;3;0].

In fig. 3.2.2 sono riportati gli andamenti del fitness medio e delle frequenze relative degli aplotipi, degli omozigoti e degli eterozigoti portatori della mutazione, a confronto tra una simulazione di selezione su carattere completamente dominante (fitness [1;1;0]) e una analoga ma su carattere completamente recessivo (fitness [1;0;0]). In queste simulazioni la mutazione compare alla generazione t=0. Si può osservare come, per la selezione su carattere recessivo (destra), la frequenza degli aplotipi mutati insieme con quella degli omozigoti mutati raggiunge rapidamente l'unità. Di fatto nelle simulazioni dopo appena 35 generazioni dalla comparsa della mutazione si ha la completa fissazione del carattere e, parallelamente, di tutti gli alleli del microsatelliti che costituiscono al sequenza di DNA degli individui.

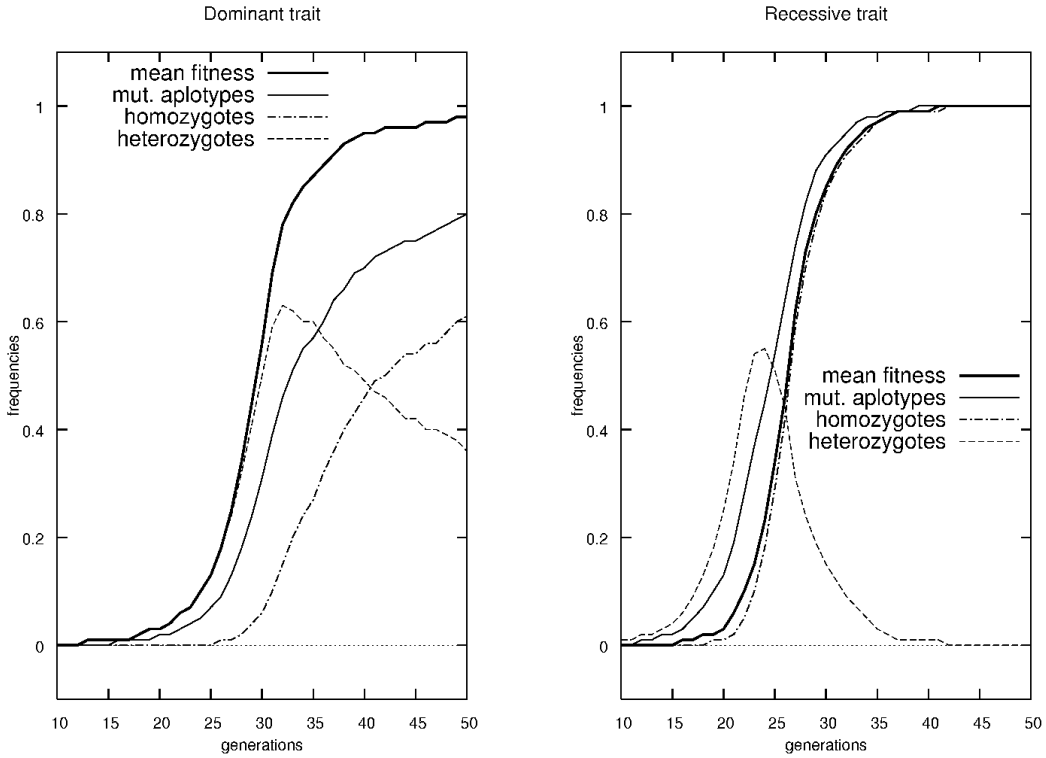


Fig. 3.2.2 confronto tra selezione su carattere dominante (sinistra) e recessivo (destra)

Il presente lavoro è volto essenzialmente allo studio dell'azione degli eventi di ricombinazione nel rilassamento del linkage disequilibrium, essendo quindi interessati all'evoluzione della popolazione nella fase immediatamente seguente lo stress selettivo, si

è preferito utilizzare un modello di selezione su carattere dominante senza che questo alterasse le conclusioni sul comportamento della popolazione in esame.

Per la popolazione studiata, tra le molte simulate, il termine del warm-up segna l'inizio della scala temporale (fig. 3.2.3). Successivamente e fino al termine della simulazione il *mutation-rate* per i microsatelliti è posto uguale a zero, giacché sulla scala temporale che si vuole osservare sono completamente trascurabili gli eventi di mutazione dei microsatelliti (tipicamente $\mu \sim 10^{-6} \div 10^{-4}$).

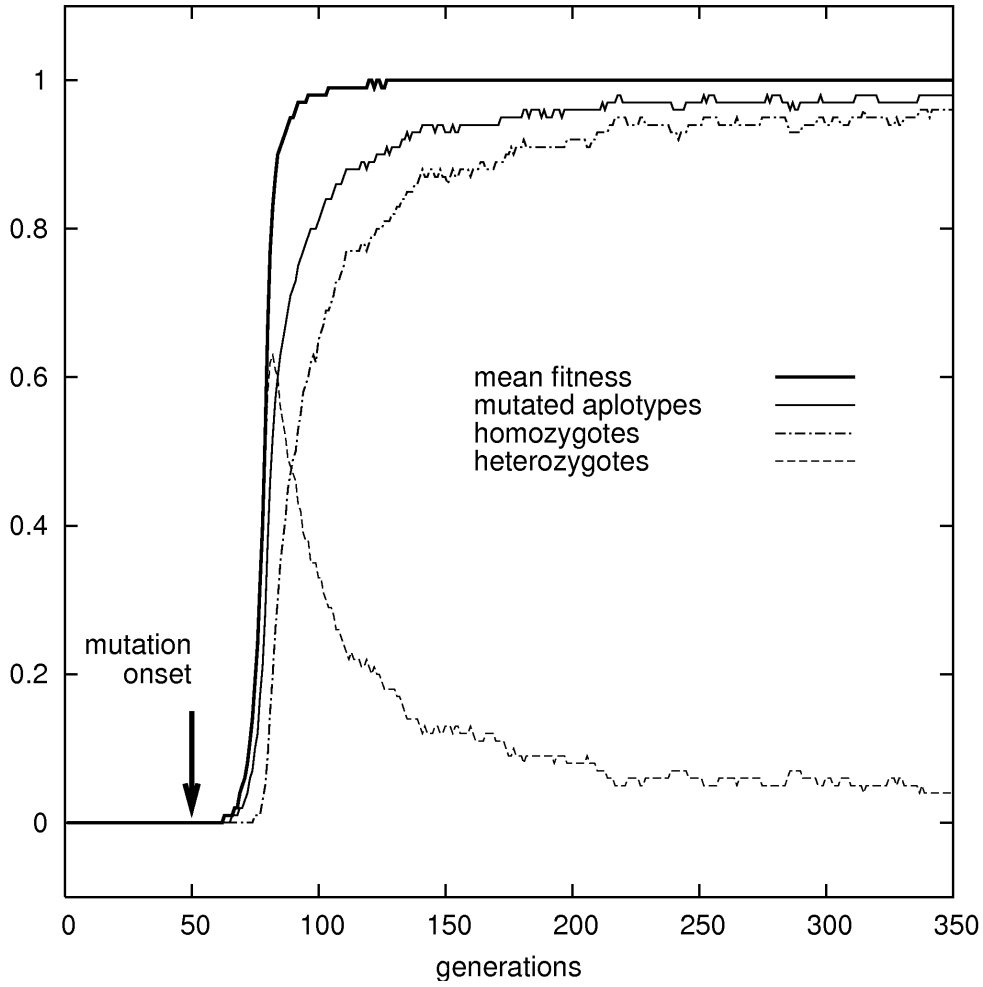


Fig. 3.2.3 diffusione della mutazione nella popolazione simulata

Dopo un periodo di stabilizzazione di 50 generazioni compare la mutazione nel locus associato al fitness ed inizia il processo di selezione sugli individui.

Il fitness medio della popolazione (linea continua più spessa) sale rapidamente fino a valori prossimi all'unità, come già descritto nel caso di selezione dominante.

Contemporaneamente la mutazione (linea continua sottile) si diffonde nella popolazione e, con un breve ritardo, anche il numero di omozigoti mutati (linea tratto-punto) cresce. Non si ha, in questo caso come nelle altre analoghe simulazioni condotte, la fissazione della mutazione nella popolazione. Rimane sempre all'interno della popolazione una percentuale di individui eterozigoti rispetto alla mutazione e ai marcatori vicini che garantisce variabilità allelica utile per la ricombinazione.

3.3 Analisi del *linkage disequilibrium*

L'indice di *linkage disequilibrium* (LD) utilizzato è il D' di Lewontin, presentato nel paragrafo 2.1, e calcolato per i microsatelliti da M1 a M9 rispetto al microsatellite M0 (completamente in link con il locus sotto selezione).

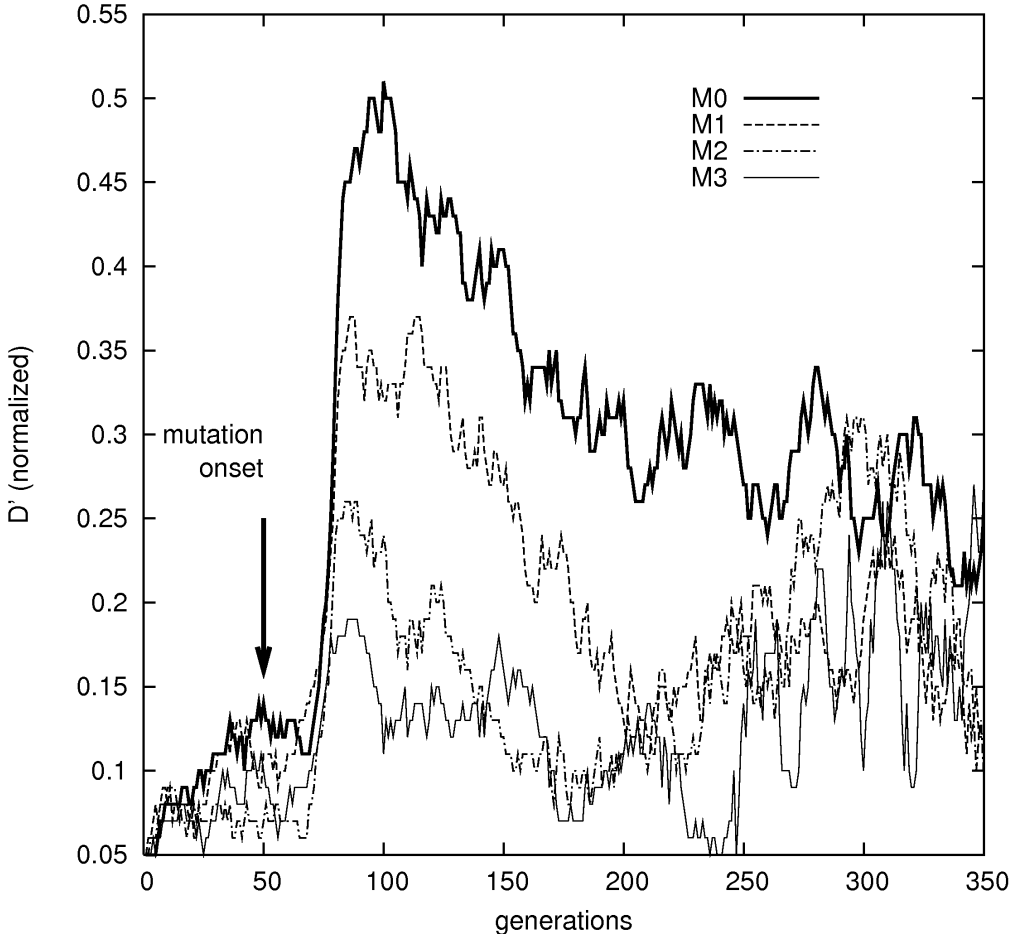


Fig. 3.3.1: andamento del LD durante la simulazione

Il comportamento di questo indice durante la simulazione è rappresentato in figura 3.3.1. L'andamento è analogo al fitness (cfr. fig. 3.2.3) fino a raggiungere valori prossimi al suo massimo intorno alla generazione t=100.

A questo punto gli effetti di selezione sulla popolazione sono trascurabili (circa l'80 percento degli aplotipi contengono l'allele mutato) e la ricombinazione tende a rilassare il LD nelle successive generazioni.

L'entità del LD indotto dalla selezione chiaramente è inversamente correlata con la distanza dei marcatori dal locus di selezione e solo i microsatelliti M1, M2 e M3 mostrano un consistente discostamento dalle condizioni di equilibrio con M0.

Dal momento che la simulazione comprende una popolazione di dimensione finita gli eventi di ricombinazione non ristabiliscono i valori di linkage disequilibrium analoghi a prima della comparsa della mutazione.

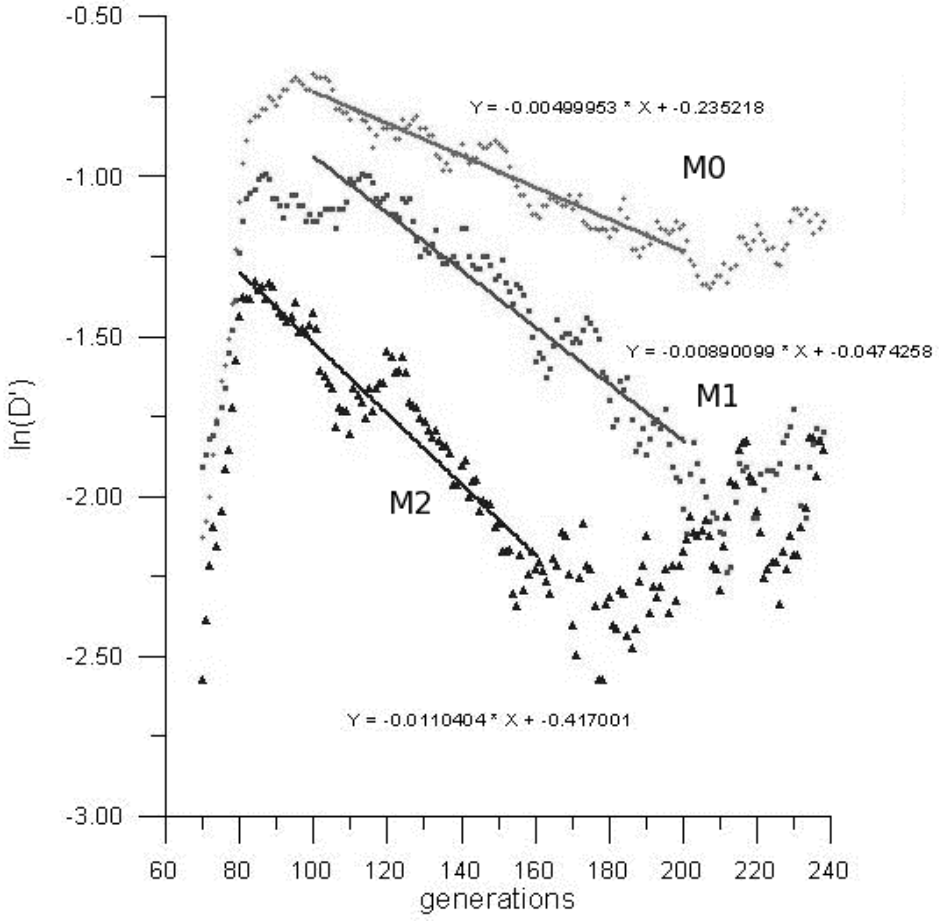


Fig. 3.3.2: *rate* di decadimento del LD durante la simulazione

Inoltre anche la rapidità del decadimento esponenziale di D' osservato è sensibilmente inferiore a quella prevista da un semplice modello di decadimento esponenziale, secondo la legge

$$D'(n)=(1-r)*D'(n-1)$$

dove n è la generazione e r e il *recombination rate* ovvero la distanza espressa in cM dal microsatellite M0.

Tuttavia l'interpolazione esponenziale (Fig. 3.3.2) nell'intervallo temporale g=(100,200) registra il giusto rapporto del *rate* di decadimento per i marcatori M1 e M2.

Durante la simulazione sono stati "raccolti" campioni, ciascuno di 50 individui, dalla popolazione alle generazioni g=50, 90, 110, 130, 150, 250, il primo dei quali alla comparsa della mutazione e i successivi durante il decadimento del linkage disequilibrium.

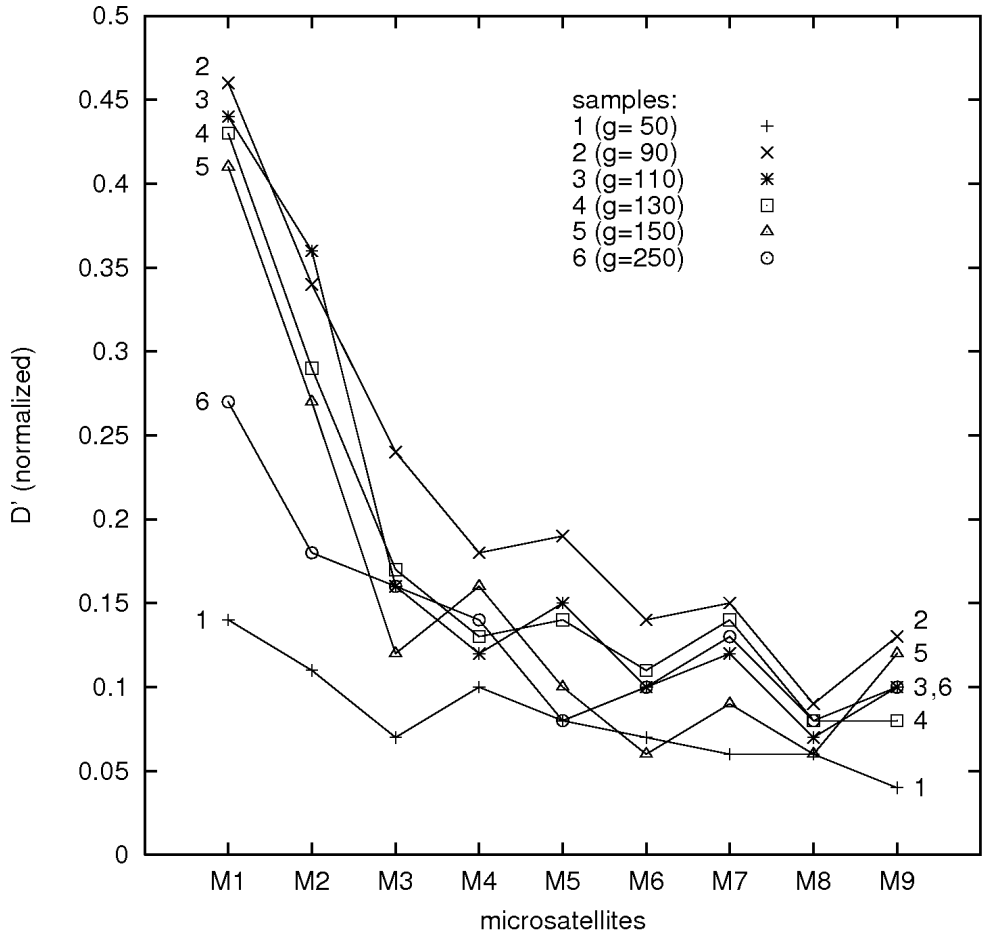


Fig. 3.3.3: *rate* di decadimento del LD durante la simulazione

Non è stato incluso nei campioni lo stato del locus sotto selezione come nella situazione tipica della ricerca di un sito sotto selezione.

La fig. 3.3.3 mostra l'indice di LD dei campioni in funzione della posizione dei microsatelliti. Alla comparsa della mutazione, nel campione 1 (g=50), gli effetti di LD sono deboli in quanto ancora assente la selezione e quello che si vede deve attribuirsi alle dinamiche della popolazione di dimensione finita.

Il campione 2 a g=90 presenta invece il massimo per l'indice di linkage disequilibrium. I campioni successivi registrano l'azione progressiva della ricombinazione e il conseguente ristabilirsi delle condizioni di equilibrio.

Comunque in tutti i campioni si notano sensibili effetti di linkage con il microsatellite M0 solo fino al marcatore M5 mostrando che in questo modello gli effetti della selezione si estendono fino a circa 10 cM dal sito di selezione.

3.4 Analisi con un modello a coalescente

Per l'analisi con un modello a coalescente si sono considerate separatamente la regione "reg1", vicina al sito di selezione e costituita dai microsatelliti M0-M4, e la regione "reg2", costituita dai rimanenti microsatelliti M5-M9.

Per queste regioni si è stimato il *rescaled mutation rate* Θ, per ciascuno dei sei campioni e per ciascuna regione indipendentemente, per mezzo dell'analisi di massima verosimiglianza eseguita utilizzando il software LAMARC 1.2.2 già citato nel paragrafo 2.3. Lungo tempo è stato dedicato alla messa a punto dei parametri di ricerca come i *warm-up* e la durata delle catene MH-MCMC, il loro numero e la loro "temperatura". Al fine di essere sicuri della convergenza dei *run* sono state ripetute più volte le stime dei Θ (fino a 20 determinazioni per ogni regione e per ogni popolazione).

In fig. 3.4.1 è mostrato il rapporto $\Theta_{reg1}/\Theta_{reg2}$ delle due regioni nei diversi campioni prelevati dalla popolazione simulata.

Si può osservare come il valore prossimo all'unità del primo campione (g=50) scenda rapidamente e raggiunga il suo minimo nel campione 3 (g=110), circa dieci generazioni dopo il massimo nel linkage disequilibrium. I valori del rapporto per i campioni successivi rimangono definitivamente al di sotto del valore 0.5.

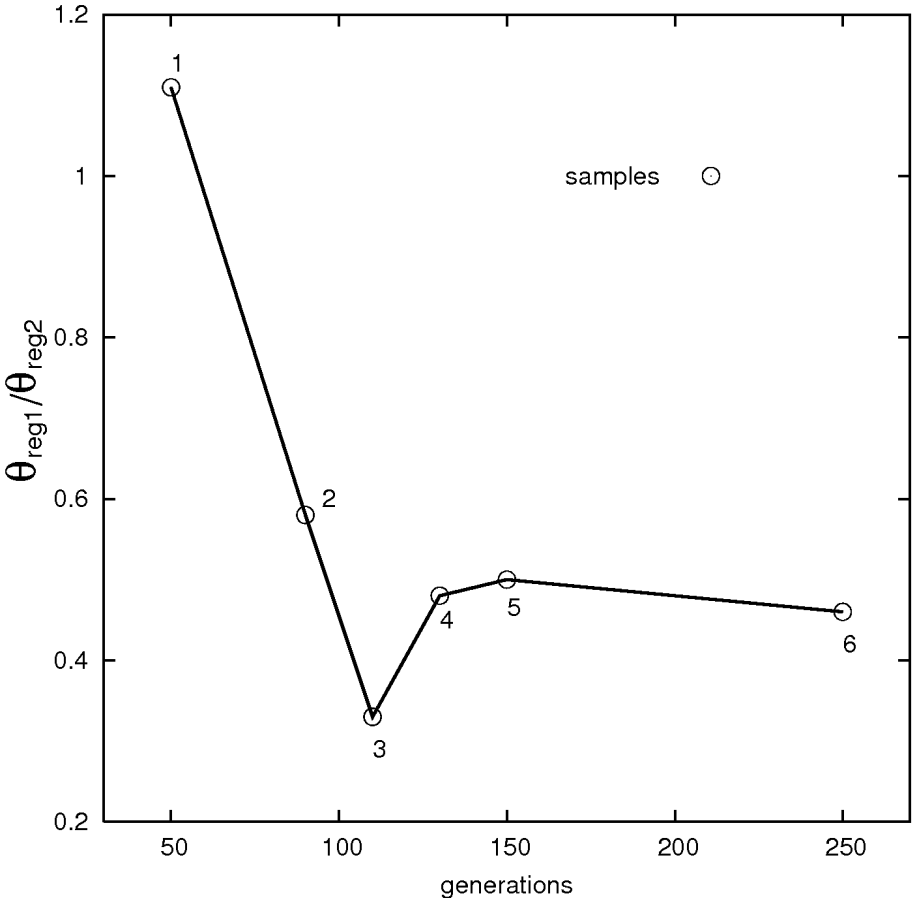


Fig. 3.4.1: Θ_{reg1} /Θ_{reg2} per i campioni durante la simulazione

Questo comportamento risulta del tutto generale per le simulazioni condotte e sempre si è registrato un abbattimento del *rescaled mutation rate* per marcatori prossimi ad un sito sotto selezione.

In figura 3.4.2 è riportato il risultato di due serie di simulazioni, la prima per dieci popolazioni di 4000 individui sotto selezione e la seconda per dieci popolazioni analoghe ma in condizioni "normali". I marcatori considerati sono anche in questo caso dieci microsatelliti disposti però in successione con cadenza di 1 cM.

I campioni, per un numero totale pari a 300 e costituiti da cinquanta individui ciascuno, sono stati raccolti con una elevata frequenza (5 generazioni) durante il massimo di stress selettivo e in modo più diradato (20 generazioni) a selezione oramai completa.

Il grafico di sinistra mostra gli andamenti di Θ per le due serie di simulazioni mentre il grafico i destra mostra i rapporti dello stesso parametro riconducibili al rapporto dei μ per

le serie di microsatelliti simulate. Le barre di errore rappresentano una indicazione della variabilità delle determinazioni tra le simulazioni, ottenuta come deviazione standard su ciascun gruppo di dieci campioni.

Anche in questo caso non si ha il rilassamento dell'indice verso le condizioni iniziali come nel caso del *linkage disequilibrium*. In effetti, l'azione della pressione selettiva riduce definitivamente la variabilità allelica nella regione vicina al sito di selezione e questa condizione può essere risolta solo attraverso eventi di mutazione e comparsa di nuovi alleli.

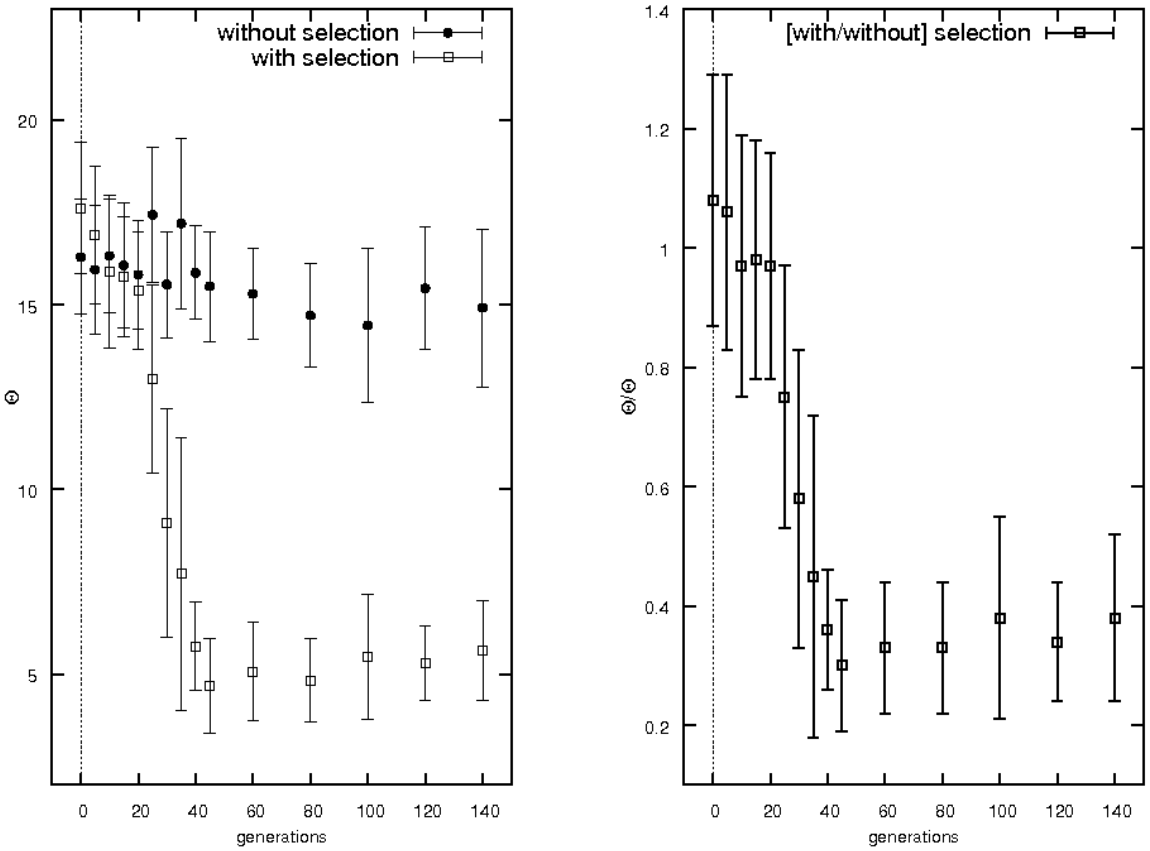


Fig. 3.4.2: Andamento di Θ per dieci campioni simulati di popolazioni sotto selezione e dieci normali (sinistra) e andamento di Θ_{sel} /Θ_{nor} per gli stessi campioni (destra). Le barre di errore indicano le deviazioni standard sulle simulazioni

4

Studio di un caso reale

In questo capitolo verrà affrontato il problema della ricerca di indicatori di selezione tramite marcatori neutrali in un caso di popolazioni reali.

Saranno applicate al caso di studio le stesse tecniche statistiche utilizzate per analizzare l'influenza di una pressione selettiva sulla popolazione simulata del capitolo precedente.

Oltre alle analisi di *linkage disequilibrium* e del *rescaled mutation rate*, avendo in questo caso campioni provenienti da più popolazioni differenti, verrà valutato anche il comportamento dell'indice di fissazione Fst considerando diverse super-popolazioni costituite raggruppando in vario modo i campioni delle popolazioni reali.

4.1 Il processo di selezione sul gene della miostatina

Si considera la selezione su mutazioni nel gene della miostatina in tre razze bovine: Belgian Blue, Piemontese e Marchigiana. Queste razze portano tre mutazioni differenti ma tutte causa di perdita di funzionalità del gene.

Il guadagno produttivo di questa mutazione e le maggiori qualità alimentari del prodotto (basso contenuto di grasso e "tenerezza") hanno prevalso sui costi e problemi di allevamento associate a questa mutazione. Infatti le razze Belgian Blue e Piemontese nel secolo scorso sono state soggette a una sistematica selezione che ha portato alla quasi completa fissazione di questa mutazione in molte popolazioni.

Il fenotipo "doppia coscia" sembra comparso nella razza Belgian Blue a metà del diciannovesimo secolo. Questa razza fu successivamente sottoposta a forte selezione specialmente negli anni '50 del ventesimo (secondo dopoguerra).

Il primo capo di razza Piemontese che presentava il fenotipo "doppia coscia" è registrato nell'*herd-book* nel 1886. Tuttavia fino agli anni 80 del secolo scorso la selezione

su questa razza dedicata alla produzione di carne non è stata molto intensa a causa dell'interesse commerciale del latte. Tuttavia negli ultimi venti anni la produzione di carne ha prevalso su quella lattiera.

Nella razza Marchigiana la mutazione del gene della miostatina non è ancora molto diffusa [16] e gli omozigoti portatori della mutazione risultano essere appena il 5% della intera popolazione. Questo induce a pensare che il processo di selezione per questa razza sia appena agli inizi

4.2 Il *dataset*

Il dataset utilizzato nel presente lavoro [17] consiste di sei campioni di individui provenienti sia da razze in cui sono presenti le mutazioni nel gene della miostatina sia da razze normali di confronto. In dettaglio:

nome	*individui*	*abbreviazione*	*mutazione*
Marchigiana	25	MC	presente
Piemontese	14	PI	presente
Belgian Blue	17	BB	presente
Chianina	16	CH	assente
Holstein	13	HO	assente
Romagnola	14	RM	assente

I marcatori analizzati sono cinque microsatelliti riportati, insieme alla loro distanza dal gene della miostatina, nella seguente tabella:

Microsatellite	Distanza dal gene della miostatina (cM)
tgla44	2
bm3627	5,6
tgla431	9
bm3010	35,3
bm440	70

Come si vede i primi tre microsatelliti si trovano a stretto contatto con il locus di mutazione, mentre gli altri due sono a distanza considerevole. Ci si aspetta quindi una situazione di forte *linkage* per i primi (in particolare tgla44) e di relativa indipendenza per i due più lontani.

Il campione della razza Marchigiana non è rappresentativo della popolazione rispetto alla mutazione del gene della miostatina, in quanto mostra un sovra-campionamento di individui mutati rispetto ai dati sulla popolazione riportati in letteratura. Il motivo è che queste misure sono state raccolte nell'ambito di uno studio sulla morfologia a livello di genetica molecolare delle mutazioni a carico del gene della miostatina . In tale lavoro si era quindi particolarmente interessati a raccogliere campioni di individui portatori della mutazione senza curarsi della rappresentatività statistica del campione stesso.

La situazione è descritta dalla seguente tabella:

	Individui del campione	*Freq. Pop. (%) in letteratura*
Omozigoti normali	7	69
Omozigoti mutati	7	5
Eterozigoti	11	26

Si può notare come la frequenza nel campione degli omozigoti portatori della mutazione (28%) sia ben lontana del 5% stimato nella popolazione reale.

Viceversa nel tipo di analisi statistiche utilizzate nel presente lavoro è della massima importanza la rappresentatività del campione.

Per cercare di ovviare a questo problema per la razza Marchigiana si è provveduto a rigenerare quattro nuovi campioni (MC1, MC2, MC3 and MC4) ottenuti tramite *bootstrap* pesato con le frequenze stimate nella popolazione reale.

Sul totale di dieci campioni (originali e rigenerati) si è provveduto a fare qualche semplice analisi statistica per essere sicuri che i campioni ottenuti da *bootstrap* fossero compatibili con le misure originali.

In fig. 4.2.1 è rappresentata una analisi di raggruppamento (UPGMA) per i campioni. La distanza utilizzata è quella minima di Nei e si vede come i campioni rigenerati formino correttamente un unica clade con il campione originale di MC (nel grafico MR). Utilizzando la stessa matrice di distanza, nel grafico di fig. 4.2.2 vi sono le prime due componenti principali che mostrano come anche in questa analisi i campioni rigenerati si collochino correttamente i prossimità del campione originale.

Evidentemente il *bootstrap* ha comunque prodotto un campione meno “estremo” di quello originale.

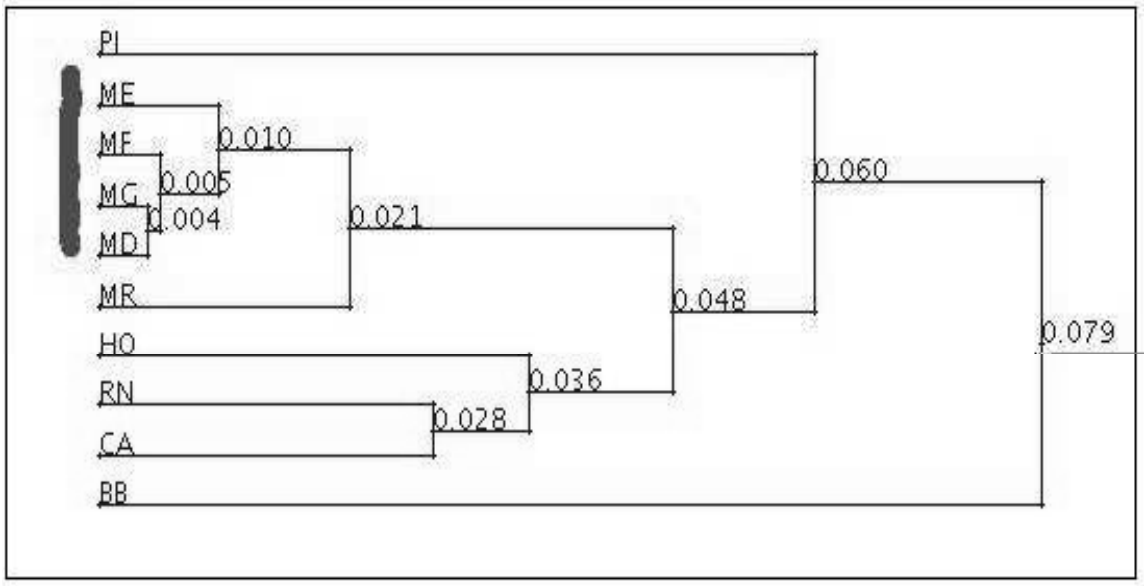


Fig. 4.2.1: UPGMA dei campioni, inclusi quelli rigenerati

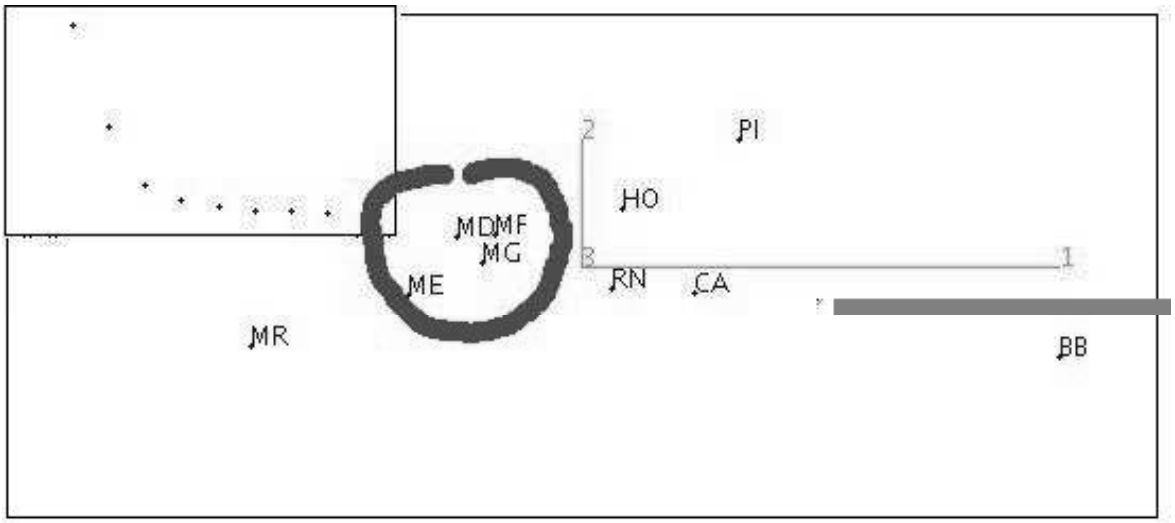


Fig. 4.2.2: prime due componenti principali per i campioni

4.3 Analisi di *linkage disequilibrium*

Per l'analisi di *linkage disequilibrium* (LD) si è utilizzato l'indice D', presentato nel paragrafo 2.1, calcolato tra il TGLA44 e i restanti microsatelliti. Sulla scala temporale del presente lavoro si può infatti considerare il TGLA44 in *linkage* completo con il gene della miostatina. I risultati per i campioni originali sono riportati in fig. 4.3.1.

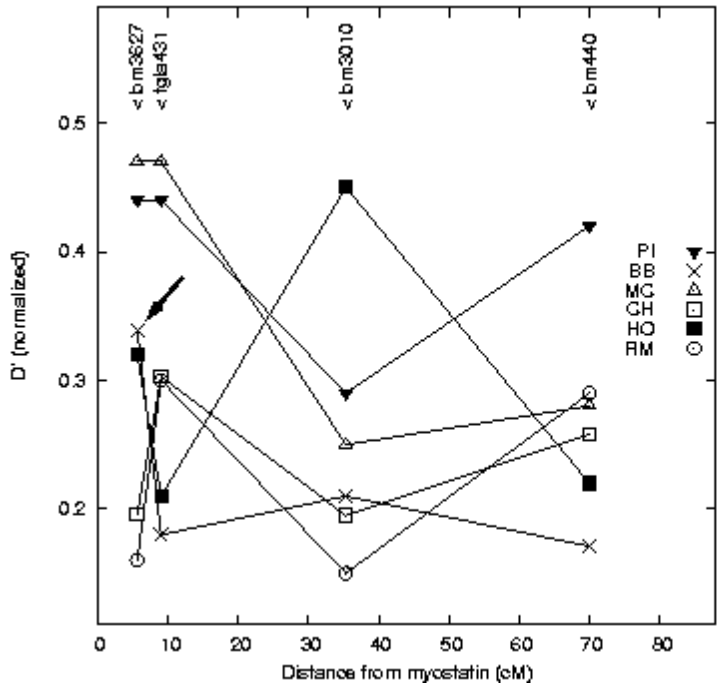


Fig. 4.3.1: *Linkage Disequilibrium* per i campioni originali

Il grafico a prima vista risulta di difficile lettura per la "rumorosità" dei dati dovuta al piccolo numero di microsatelliti e alla limitatezza dei campioni utilizzati. Tuttavia alcune informazioni possono essere desunte per i due microsatelliti BM3627 e TGLA431 prossimi al sito di selezione.

Per i campioni originali delle razze Piemontese (triangoli pieni) e Marchigiana (triangoli vuoti) il D' tra questi microsatelliti e il TGLA44 risulta effettivamente molto alto rispetto a quello delle razze "normali" Chianina, Holstein e Romagnola (quadrati e cerchio) indicando quindi un evidenza di possibili effetti di selezione sulle due razze.

Per la razza Belgian Blue (freccetta sul BM3627) viceversa non sembra essere molto evidente alcun effetto di selezione, almeno in grado di distinguersi dalle fluttuazioni delle

razze normali.

Questi risultati sembrerebbero in contraddizione con il dato storico che, ricordiamo, documenta una intensa attività di selezione sulle razze Belgian Blue e Piemontese e scarsa attività sulla razza Marchigiana.

Per quest'ultima razza l'incongruenza del risultato sul LD è chiarita dalla stessa analisi eseguita sui campioni ricostruiti e maggiormente rappresentativi della popolazione reale.

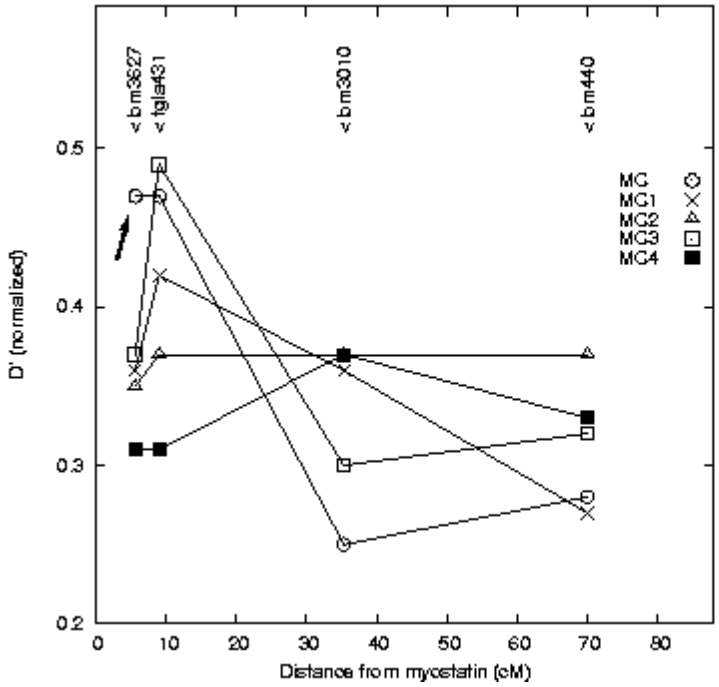


Fig. 4.3.2: *Linkage Disequilibrium* per i campioni della razza Marchigiana

In fig. 4.3.2 è riportato lo stesso D' di fig. 4.3.1 ma questa volta per il campione originale della razza Marchigiana e per i quattro campioni ottenuti via *bootstrap* (in questo grafico indicati con MC 1,2,3 e 4). Come si vede il LD tra il TGLA44 e il BM3627 del campione originale (freccetta) è consistentemente ridimensionato nei campioni rigenerati.

In conclusione, in base a questo tipo di analisi, possiamo concludere che solo il campione della razza Piemontese mostra un consistente scostamento del D' dai valori aspettati in condizione normali. Quindi solo per questa razza possiamo confermare la presenza di una pressione selettiva sul gene della miostatina.

Le indicazioni sul campione originale della razza Marchigiana derivano si da un processo di selezione ma avvenuto non sulla popolazione bensì nella fase di raccolta del campione con l'eccesso di individui portatori della mutazione.

4.4 Analisi dell'indice di fissazione Fst

I valori dell'indice di fissazione Fst ottenuti dai rapporti di eterozigosità per i campioni osservati e per i campioni della razza Marchigiana rigenerati tramite *bootstrap* sono presentati in fig. 4.4.1.

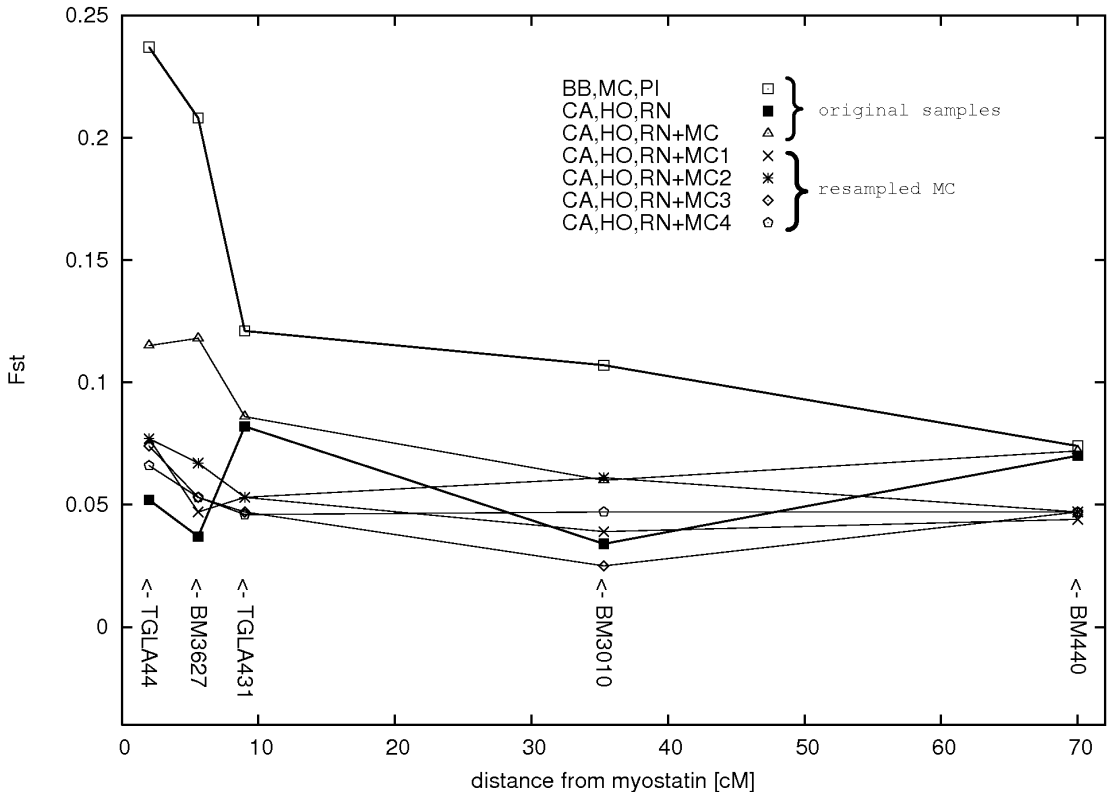


Fig. 4.4.1: Comportamento dell'indice Fst in corrispondenza del microsatelliti osservati

Per ciascun microsatellite è riportato il valore di tale indice considerando ciascuna razza come sottopopolazione di diversi possibili raggruppamenti delle popolazioni originali. Il valore puntuale dell'indice Fst in ciascuna razza risulta di difficile lettura ancora una volta a causa della esiguità dei campioni e del numero di marcatori considerati. Per questo

motivo ciascun dato rappresentato nel grafico è ottenuto mediando il valore di Fst dei diversi campioni considerati nel raggruppamento.

Per il raggruppamento delle popolazioni in cui è presente la mutazione (quadratino vuoto) si nota come il valore dell'Fst sia molto elevato in corrispondenza dei microsatelliti TGLA44 e BM3627 prossimi al sito di selezione.

Al contrario per le popolazioni "normali" (quadratino pieno) il valore dell'Fst risulta più basso per tali marcatori e consistente con quello dei microsatelliti più lontani (BM3010 e BM440) dal gene della miostatina.

È interessante osservare i risultati per il raggruppamento di popolazioni ottenuto dalle popolazioni normali aggiungendo il campione originale della razza Marchigiana (triangoli vuoti). L'aggiunta di questa popolazione innalza sensibilmente il valore dell'Fst di questo raggruppamento ancora una volta in corrispondenza dei marcatori prossimi al sito di selezione.

La stessa cosa non accade se si introducono nel raggruppamento delle razze "normali" i diversi campioni ottenuti da *bootstrap* e che sono quindi rappresentativi dell'attuale stato di diffusione della mutazione nella popolazione della razza Marchigiana.

Come nel caso del analisi di *linkage disequilibrium* si nota la criticità del metodo rispetto alla qualità dei campioni in esame.

Meno evidente l'effetto del campionamento nell'analisi dell'indice Fst condotta confrontando il dato osservato con una simulazione di un modello teorico "a isole", già descritto nel paragrafo 2.2.

In figura 4.4.2 sono riportati quattro grafici con il posizionamento dei microsatelliti in diversi raggruppamenti di popolazioni sulla distribuzione teorica dei valori dell'indice di Fst e nella banda di confidenza stimata dalla simulazione.

La linea continua rappresenta il valore aspettato per l'Fst del modello e le linee tratteggiate la banda di confidenza al 5%.

I due grafici di sinistra si riferiscono rispettivamente al raggruppamento delle razze "normali" (grafico superiore) e al raggruppamento delle razze portatrici di mutazioni sul gene della miostatina (grafico inferiore). Si può osservare come nel caso delle razze mutate i microsatelliti TGLA44 e BM3627 si trovino al di fuori della banda di confidenza. I due grafici di destra si riferiscono invece all'introduzione nel raggruppamento delle razze "normali" del campione originale della razza Marchigiana (grafico superiore) e della razza

Piemontese (grafico inferiore). Nel caso dell'introduzione della razza Piemontese si nota come il microsatellite TGLA44 si trovi al di fuori della banda di confidenza segnalando quindi l'anomalia introdotta in questo gruppo di popolazioni. Viceversa il campione originale della razza Marchigiana non altera in modo così consistente il comportamento dell'Fst.

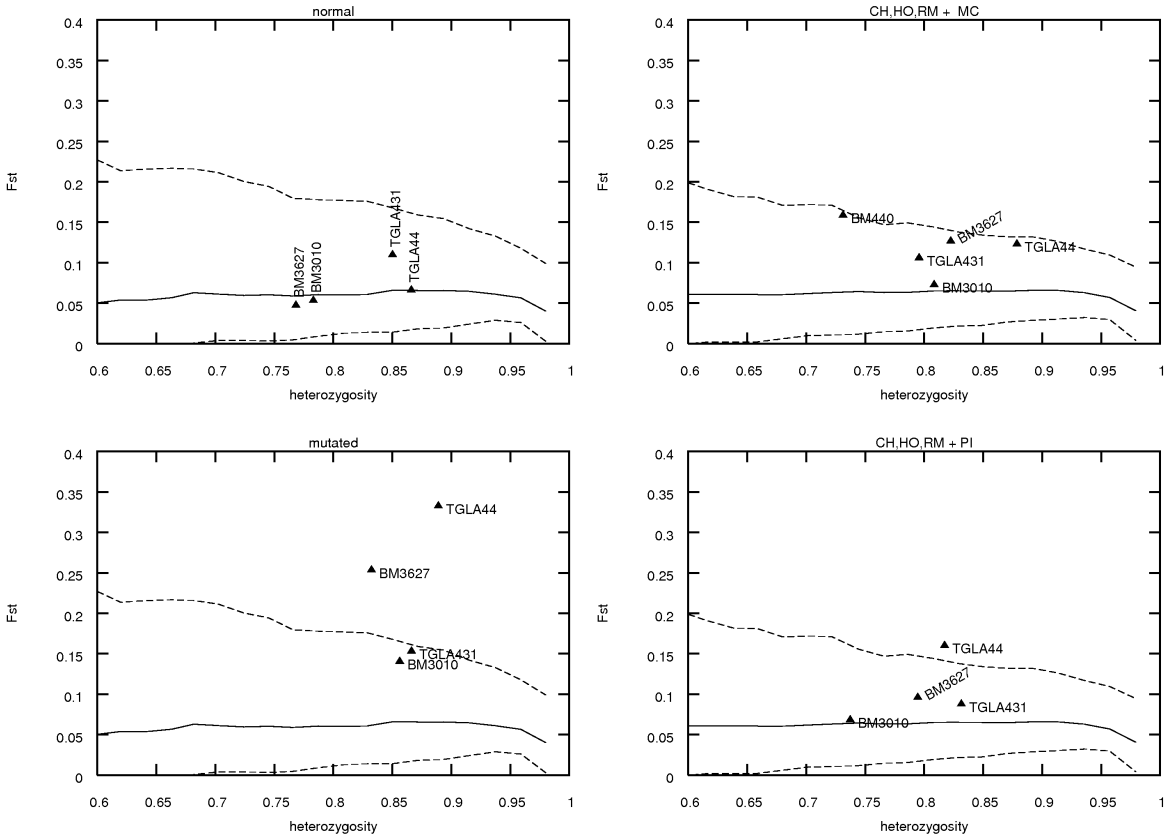


Fig. 4.4.2: Ricerca di marcatori *outliers* in un modello "a isole": campioni originali

Anche se le determinazioni di questo indice per tre dei microsatelliti si trovano al limite della banda di confidenza del 5 percento, debbono comunque essere considerate fluttuazioni casuali accettando l'ipotesi nulla di nessuna selezione.

Si è condotto lo stesso tipo di analisi introducendo questa volta i campioni ricostruiti per la razza Marchigiana all'interno del raggruppamento delle razze "normali". Il risultato è mostrato in figura 4.4.3.

In questo caso i quattro microsatelliti si collocano ben all'interno della banda di confidenza in prossimità del valore dell'indice di Fst aspettato.

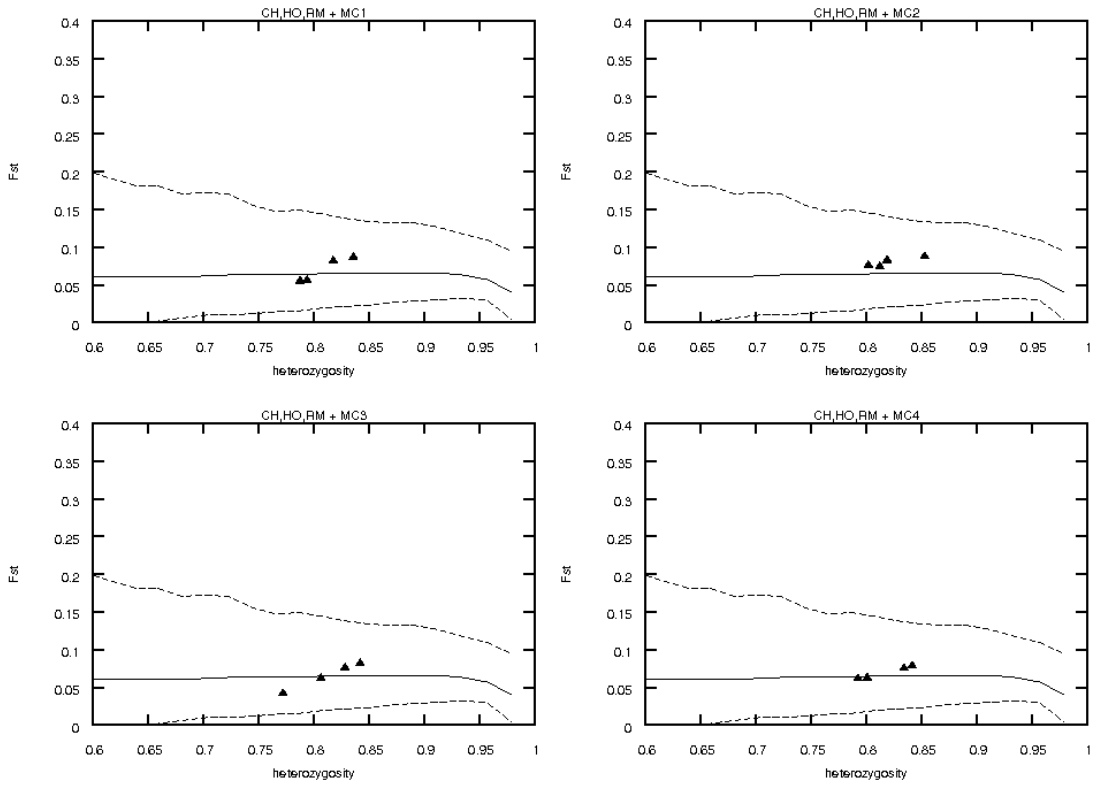


Fig. 4.4.3: Ricerca di marcatori *outliers* in un modello "a isole": campioni ricostruiti

4.5 Analisi con un modello a coalescente

Al fine di osservare gli effetti di selezione con un modello a coalescente, come nel caso della popolazione simulata, si sono divisi i marcatori in due gruppi: i tre più vicini (TGLA44, BM3627, TGLA431) denominati "reg1" e i restanti due (BM3010, BM440) denominati "reg2".

Per entrambi i gruppi e per ciascuna popolazione è stato stimato il *rescaled mutation rate* Θ e in fig. 4.5.1 è mostrato il rapporto tra questo parametro nelle due regioni per ciascun campione di razza.

Per la Romagnola il risultato mostrato non è di nessuna indicazione in quanto, pur allungando notevolmente il numero di passi delle catene di Markov del software non si è riusciti ad ottenere la riproducibilità del dato nei diversi *run*.

Rimangono quindi come campioni di riferimento solamente quelli provenienti dalle razze Holstein e Chianina.

Considerando questi come rappresentativi di razze "normali" dobbiamo assumere che il valore del rapporto intorno a 0.3 indichi effettivamente un diverso valore dei Θ tra i due *set* di microsatelliti anche in condizione di assenza di selezione.

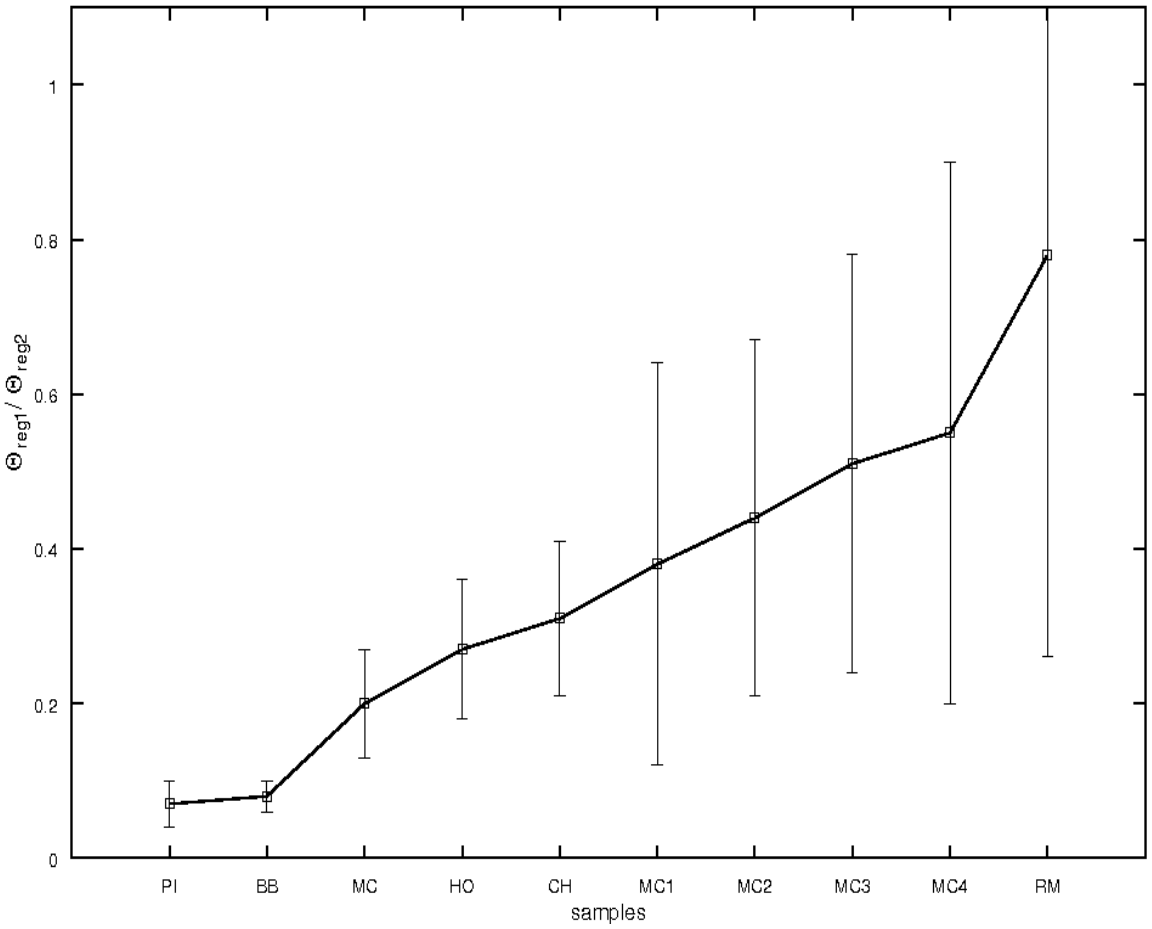


Fig. 4.5.1: $\Theta_{reg1}/\Theta_{reg2}$ per i campioni reali e ricostruiti

Lo stesso rapporto per le razze Piemontese e Belgian Blue risulta apprezzabilmente inferiore indicando valori di Θ più bassi per il *set* di microsatelliti vicini al gene della miostatina. Questo è assolutamente compatibile con un processo di selezione che può essere visto come un agente che riduce la variabilità allelica dei marcatori vicini al sito di selezione, come si è visto anche nel caso della popolazione simulata.

In altre parole per i marcatori della "reg1" la lunghezza complessiva dell'albero di coalescenza stimato risulta in queste razze molto più breve di quella stimata per i marcatori della "reg2".

Il campione originale della razza Marchigiana viene correttamente descritto più vicino

alle razze “normali” che non alla razze sotto selezione. Questo risultato indica come il coalescente sia più legato alla tipologia degli alleli che non alle loro frequenze e, contrariamente a quanto accade per il LD, non si lasci ingannare da un cattivo campionamento.

Per i campioni ricostruiti per la razza Marchigiana l'analisi di questo grafico sembra risultare di difficile interpretazione. La spiegazione di questo sarà possibile averla analizzando il rapporto dei Θ tra le popolazioni.

La fig. 4.5.2 rappresenta il grafico dei rapporti di Θ rispetto a quello del campione della razza Holstein nelle due regioni. Chiaramente per questa razza tale rapporto è pari all'unità.

Si può osservare come anche per le razze Chianina e Marchigiana (campione originale) non si abbiano valori molto diversi dall'unità e soprattutto tali valori siano molto simili nelle due regioni. Anche questo indice descrive sulla razza Marchigiana lo scarso effetto della pressione selettiva sulla miostatina.

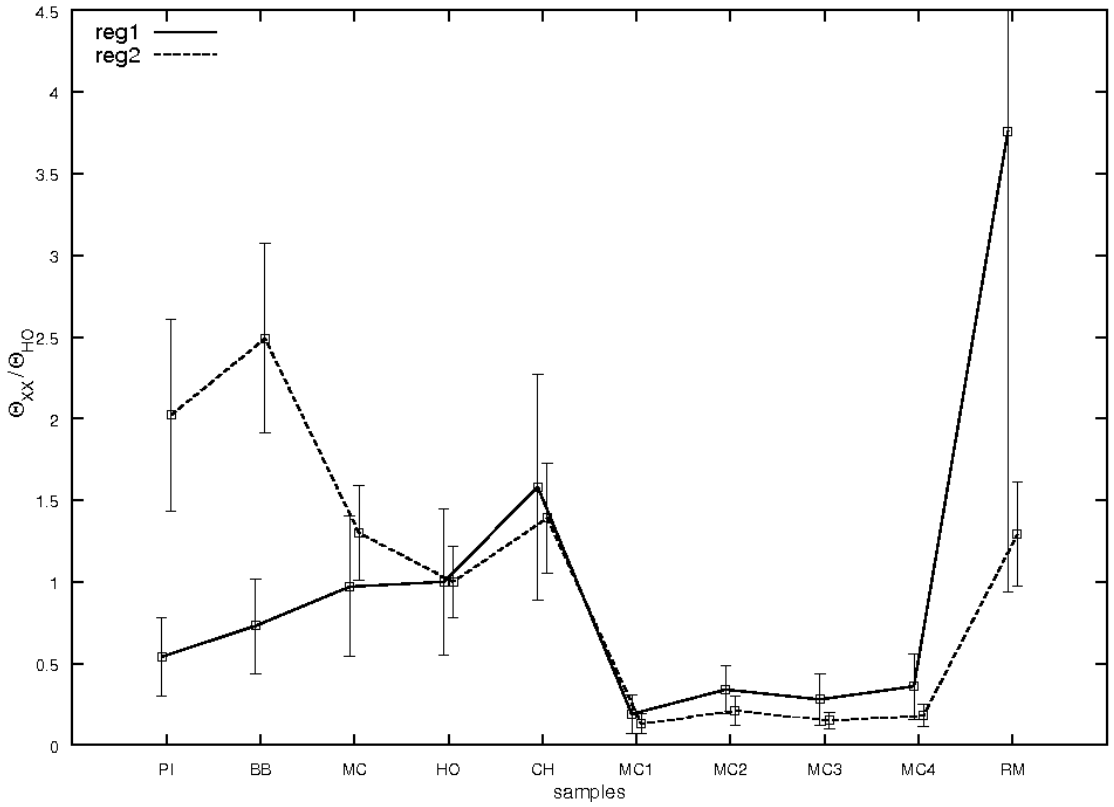


Fig. 4.5.2: Θ_{XX}/Θ_{HO} nelle regioni “reg1” e “reg2”

Come si è accennato nel paragrafo 2.3 il rapporto di Θ tra popolazioni differenti è legato alla stima del rapporto delle dimensioni delle popolazioni effettive per i diversi campioni. In questo caso, per le razze sotto selezione, sembrerebbe descrivere un aumento di tali dimensioni nella regione "reg2" rispetto alle razze "normali".

Tuttavia bisogna sottolineare come per le popolazioni allevate artificialmente sia estremamente difficile, quando non rischioso, definire una popolazione effettiva e quindi non sembra possibile dare una interpretazione assoluta dei valori riscontrati. Resta però ancora molto evidente come un processo di selezione alteri sostanzialmente il risultato di una stima attraverso un modello a coalescente.

Sempre senza voler attribuire numeri concreti a questa analisi possiamo però ottenere una indicazione sulla anomala stima ottenuta per i campioni ricostruiti per la razza Marchigiana nel precedente grafico di fig. 4.5.1. Nel grafico in fig. 4.5.2 infatti si nota come il rapporto dei Θ relativo al rapporto delle dimensioni delle popolazioni effettive per i campioni ricostruiti sia sostanzialmente inferiore rispetto a quello del campione originale e anche di tutti gli altri campioni. La spiegazione si ha ricordando che tali campioni sono ottenuti come *bootstrap* dall'originale e il modello si accorge dei numerosi cloni presenti, attribuendo quindi una dimensione della popolazione effettiva molto bassa.

Si può quindi ipotizzare che la tecnica di *bootstrap* non sia compatibile con questo tipo di analisi e conseguentemente considerare privi di senso per questi campioni anche i risultati del precedente grafico di fig. 4.5.1.

5

Discussione e conclusioni

5.1 Popolazione simulata

Nel caso della popolazione simulata, sia l'analisi del linkage disequilibrium che le stime tramite il modello a coalescente hanno mostrato un considerevole discostamento dei campioni dalla condizione di equilibrio.

Mentre gli eventi di ricombinazione, con il trascorrere delle generazioni, rilassano l'indice di linkage disequilibrium verso valori normali, il parametro Θ non ritorna ai valori iniziali e questo perché tale parametro è intimamente legato alla ricchezza allelica del campione. Solo eventi casuali di mutazione dei marcatori possono porre rimedio alla riduzione del pool di alleli a disposizione di una popolazione isolata ma, come più volte ripetuto, il loro verificarsi su queste scale temporali è del tutto trascurabile.

Sia l'analisi di linkage disequilibrium che l'analisi con un modello a coalescenza si dimostrano utili negli studi di selezione, il primo per evidenziarne gli effetti recenti o attuali sulla popolazione e possibilmente datare il momento di massimo stress, il secondo per scoprire le tracce di selezione quando ormai il linkage disequilibrium non è più informativo.

5.2 Popolazioni reali

Nel caso di popolazioni reali, solo per il campione della razza Piemontese l'analisi di linkage disequilibrium indica una pressione selettiva ancora in corso o molto recente.

Al contrario, l'analisi con un modello a coalescenza sugli stessi campioni mostra correttamente un consistente effetto di selezione sia per la razza Piemontese sia per la razza

Belgian Blue, confermando contemporaneamente un ancora piccolo effetto della selezione sulla razza Marchigiana.

Il vantaggio di questo metodo risiede infatti proprio nella sua capacità di stimare la "storia" del campione e questa conserva traccia di un eventuale processo selettivo più a lungo di quanto non faccia la composizione aplotipica del suo patrimonio genetico.

5.2.1 Effetti di selezione nel campione della razza Marchigiana

I campioni riprodotti tramite bootstrap sui giusti rapporti tra aplotipi mutati e normali restituiscono valori di linkage disequilibrium quasi normali (almeno tra i marcatori più prossimi alla miostatina). L'analisi con il modello a coalescente conferma anche sul campione originale che il patrimonio allelico non ha subito ancora l'azione di una consistente pressione selettiva.

Gli apparenti effetti di selezione sul campione originale della razza Marchigiana sono quindi un involontario artefatto dovuto alla non rappresentatività del campione.

Ovvero si potrebbe affermare che l'apparente selezione sul campione non si riferisce all'opera degli allevatori ma a quella di chi ha raccolto i dati.

Sottolineo ancora una volta che questi dati provengono da uno studio sulla morfologia della mutazione del gene della miostatina nella razza Marchigiana. Non era quindi interesse di chi li ha acquisiti rispettare nel campione le proporzioni reali della popolazione.

5.2.2 Effetti di selezione nel campione della razza Belgian Blue

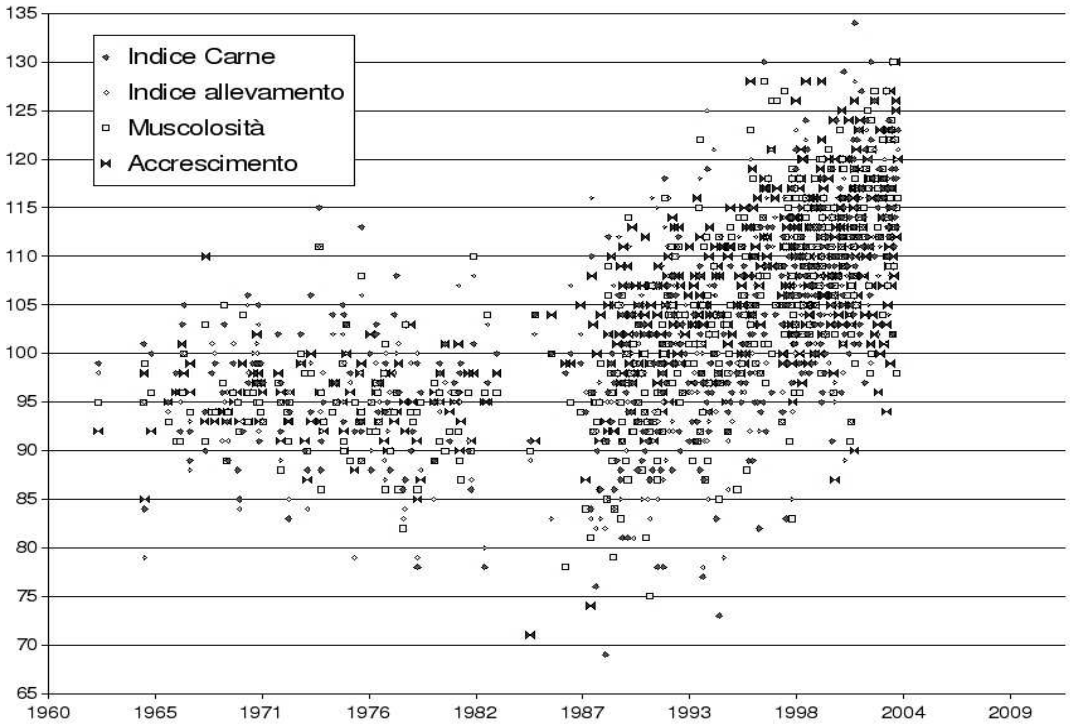
Anche nel caso della razza Belgian Blue non è osservabile un reale disequilibrio. Quella possibile indicazione è comunque indistinguibile dal "rumore" generale dei dati dovuto al limitato numero di individui e di marcatori utilizzati.

Il risultato così diverso delle due analisi sullo stesso campione della razza Belgian Blue puo' essere interpretato da un punto di vista storico: l'azione di selezione su questa razza è sufficientemente antica da permettere agli eventi di ricombinazione di ristabilire l'equilibrio tra i marcatori osservati e nel caso della razza Belgian Blue le fonti storiche riportano la parte più consistente della pressione selettiva proprio nella prima metà del novecento.

5.2.3 effetti di selezione nel campione della razza Piemontese

La razza Piemontese è l'unica che sembra avere effetti di una pressione selettiva nella regione che corrisponde al gene della miostatina.

A supporto di ciò si possono citare i dati disponibili dall'Associazione Nazionale degli Allevatori di Bovini di Razza Piemontese (ANABORAPI). Nella grafico seguente sono riportati indici di interesse zootecnico relativi alle masse muscolari di circa 1200 tori da riproduzione registrati nelle loro banche dati a partire dai primi anni '60 del secolo scorso.



E' piuttosto evidente come tutti questi indici siano stati generalmente stazionari fino alla fine degli anni '80 e stiano subendo un deciso incremento negli ultimi venti anni. Chiaramente tali indici non si riferiscono esclusivamente al gene della miostatina e al fenotipo “doppia coscia” ma confermano il recente interesse in questa razza per la produzione di carne.

5.2.4 Analisi dell’indice di fissazione Fst

Anche l’analisi dell’indice di fissazione Fst si dimostra utile ai fini della ricerca degli effetti su marcatori neutrali dovuti alla prossimità con un sito sotto selezione.

Sia la stima ottenuta dal calcolo diretto dei rapporti di eterozigosità sia quella ottenuta dal calcolo dei rapporti delle varianze, mostrano come i microsatelliti prossimi al gene mutato della miostatina siano in una condizione alterata nelle razze in cui sono presenti tali mutazioni rispetto a quanto accade nelle razze "normali".

Tuttavia anche questo tipo di analisi sembra sensibile, come l'analisi di *linkage disequilibrium*, agli effetti di cattivo campionamento per la razza Marchigiana. Entrambi i tipi di analisi presentano la debolezza di limitare il loro strumento di indagine alle frequenze rispettivamente aplotipiche il LD e alleliche l'Fst.

In particolare la strategia dell'Fst utilizzata da Beaumont non considera l'eterozigosità effettiva del campione ma solo quella aspettata in base alle frequenze alleliche. Purtroppo nel caso del presente lavoro, non è stato possibile con tale strategia riportare risultati convincenti.

5.3 Conclusioni

Dal confronto delle analisi del *linkage disequilibrium* e del *rescaled mutation rate* del modello a coalescente, sia nel caso di una popolazione simulata che nel caso di popolazioni reali, è possibile trarre principalmente due conclusioni.

Innanzi tutto la "disomogeneità" statistica degli aplotipi osservata nella analisi di *linkage disequilibrium* decade molto più rapidamente di quanto non accada per l'impronta dovuta alla selezione sulla ricchezza del *pool* di alleli della popolazione in esame. Per la disposizione dei marcatori studiati in questo lavoro mentre l'analisi di *linkage,* dopo appena qualche decina di generazioni e a causa degli eventi di ricombinazione, "dimentica" la storia di selezione cui è stata soggetta la popolazione, l'analisi utilizzando un modello a coalescente permette comunque di evidenziarne le tracce.

Inoltre il *rescaled mutation rate* stimato attraverso un modello a coalescente si dimostra meno sensibile ad un inaccurato e limitato campionamento. Tale indice risulta infatti più associato all'esistenza di un particolare allele che non alla sua distribuzione statistica all'interno di una popolazione.

Ringraziamenti

Devo innanzi tutto ringraziare per questo lavoro tutto il personale del Dipartimento di Produzioni Animali della Facoltà di Agraria dell'Università della Tuscia che in vario modo mi ha introdotto nel mondo della genetica animale.
Un ringraziamento particolare, per la loro assistenza e pazienza, la disponibilità e le vivaci discussioni, alle ricercatrici del Laboratorio di Genetica Molecolare del Prof. Alessio Valentini (in ordine *rigorosamente* alfabetico): Irene Cappuccio, Alessandra Crisà, Cinzia Marchitelli, Lorraine Pariset, Maria Carmela Savarese, Alice Valentini.
Uno speciale ringraziamento inoltre a Gabriella Porcai che continua a sopportarmi da tempo, non solo nel mio lavoro.

Bibliografia

[1] "Ex situ cryoconservation of genomes and genes of endangered cattle breeds by means of modern biotechnological methods" in FAO ANIMAL PRODUCTION AND HEALTH PAPER 76,Roma 1989 (http://www.fao.org/DOCREP/004/T0094E/T0094E01.htm)

[2] Lewontin R.C., Krakauer J. (1973)
Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms.
Genetics May;74(1):175-95

[3] Lewontin R.C. (1964)
The interaction of selection and linkage. I. General considerations; heterotic models
Genetics 49: 49-67.

[4] Hedrick P.W. (1987)
Gametic Disequilibrium Measures: Proceed With Caution
Genetics 117: 331-341

[5] Weir B.S., Cockerham C.C. (1984)
Estimating F-statistics for the analysis of population structure
Evolution 39(6): 1358-1370

[6] Nei M. (1977)
F-statistics and analysis of gene diversity in subdivided populations.
Ann. Hum. Genet. 41:225-233

[7] Beaumont M.A., Nichols R.A., (1996)
Evaluating loci for use in the genetic analysis of population structure.
Proc. R. Soc. Lond. B 263: 1619-1626.

[8] Kingman J.F. (2000)
Origins of the coalescent. 1974-1982.
Genetics Dec;156(4):1461-3

[9] Nordborg M. (2001)
in D.J. Balding, M.J. Bishop, C. Cannings (Eds)
"Handbook of Statistical Genetics"
John Wiley & Sons, Chichester, UK, 2001, p. 602

[10] Nordborg M, Tavare S. (2002)
Linkage disequilibrium: what history has to tell us.
Trends Genet. Feb;18(2):83-90

[11] Kuhner M.K., Yamato J., Beerli P., Smith L.P., Rynes E., Walkup E., Li C., Sloan J., Colacurcio P., Felsenstein J., (http://evolution.gs.washington.edu/lamarc.html)

[12] Clark A.G. (1990)
Inference of haplotypes from PCR-amplified samples of diploid populations.
Mol. Biol. Evol. Mar;7(2):111-22

[13] Excoffier L., Slatkin M. (1995)
Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.
Mol. Biol. Evol. Sep;12(5):921-7

[14] Stephens M., Donnelly P. (2003)
A comparison of bayesian methods for haplotype reconstruction from population genotype data.
Am J Hum Genet. Nov;73(5):1162-9
[15] Stephens M., Smith N.J., Donnelly P. (2001)
A new statistical method for haplotype reconstruction from population data.
Am J Hum Genet. Apr;68(4):978-89
[16] Marchitelli C., Savarese M.C., Crisa A., Nardone A., Marsan P.A., Valentini A. (2003)
Double muscling in Marchigiana beef breed is caused by a stop codon in the third exon of myostatin gene.
Mamm. Genome. Jun;14(6):392-5.
[17] Savarese M.C., Marchitelli C., Crisà A., Filippini F., Valentini A., Nardone A. (2003)
Italian Journal of Animal Science 2:64-66

Molti altri lavori per i quali non esiste un riferimento preciso all'interno di questo testo sono stati di supporto costante nel presente lavoro. Tra questi bisogna comunque ricordare:

Balding D.J., Bishop M.J., Cannings C. (Eds) "Handbook of Statistical Genetics"
John Wiley & Sons, Chichester, UK, 2001

Felsenstein J., "Theoretical Evolutionary Genetics."
(mai dato alle stampe!) http://evolution.genetics.washington.edu/pgbook/pgbook.html

Nei M., Kumar S. "Molecular Evolution and Phylogenetics"
Oxford University Press, New York, NY, 2000

Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. "Numerical Recipes in FORTRAN: the art of scientific computing"
Cambridge University Press, New York, NY, 1992

Riassunto

Metodi classici e modelli a coalescente nella ricerca di "selection signatures" tramite marcatori neutrali

La perdita di diversità biologica avviene a tutti i livelli: ecosistemi e comunità sono degradati e distrutti e molte specie sono condotte ad estinzione. Questo accade sia nelle aree tropicali che nelle zone temperate, sia negli habitat terrestri che in quelli acquatici. Anche nelle specie che comunque sopravvivono si ha una progressiva perdita di diversità genetica, da un lato per la riduzione del numero di individui che le costituiscono, dall'altro per il crescente isolamento delle popolazioni le une dalle altre.

La diversità genetica si sta perdendo anche nelle specie domestiche destinate alla produzione alimentare o comunque industriale dal momento che la produzione abbandona le tecniche di allevamento o coltivazione tradizionale in favore di metodi che tendono a replicare massivamente gli individui che corrispondono a particolari criteri commerciali su larga scala. Le nuove tecnologie rendono anche disponibili metodiche di riproduzione artificiale o comunque assistita che permette la diffusione di tipi genetici su aree geografiche oltremodo estese con la progressiva distruzione di popolazioni autoctone.

Il risultato è una considerevole pressione selettiva su queste popolazioni animali con il conseguente impoverimento della variabilità genetica e il suo appiattimento sul piano geografico.

Nella situazione attuale è quindi della massima importanza lo studio e il controllo delle dinamiche selettive al fine di prevenire danni irreparabili alla varietà genetica delle popolazioni allevate.

Obiettivo del presente lavoro è quindi quello di studiare la tracciabilità degli effetti di selezione genetica tramite l'uso di marcatori neutrali e degli strumenti statistici della genetica di popolazione.

Come caso di studio viene scelto il gene associato alla sintesi della miostatina, una proteina moderatrice dello sviluppo muscolare del feto. In alcune razze bovine varie mutazioni sono responsabili di un difetto in tale gene e quindi della carenza della proteina associata. Queste mutazioni implicano quindi uno sviluppo anormale delle masse muscolari degli

individui portatori e ciò costituisce per gli allevatori un estremo interesse dal punto di vista commerciale sia per le prestazioni produttive di questi individui sia per le caratteristiche nutrizionali delle carni prodotte.

Vengono quindi considerati sei campioni provenienti da tre razze (Piemontese, Marchigiana e Belgian Blue) sotto selezione per tali mutazioni e da tre razze "normali" utilizzate come confronto.

I marcatori genetici utilizzati sono cinque microsatelliti situati in prossimità del gene della miostatina che, ci si aspetta, dovranno evidenziare gli effetti di selezione sulle razze interessate dalle mutazioni.

Preventivamente allo studio diretto del caso delle popolazioni reali viene condotta una serie di simulazioni su calcolatore in grado di rappresentare in modo realistico le popolazioni bovine da cui provengono i campioni reali studiati nel presente lavoro. Su una di queste popolazioni simulate viene indotta la comparsa di una mutazione sulla quale si esercita una pressione selettiva favorevole analoga a quella applicata nel caso della mutazione del gene della miostatina.

La simulazione permette di osservare il comportamento del linkage disequilibrium di marcatori microsatelliti in prossimità del sito di selezione. Durante la fase più critica del processo selettivo, dalla comparsa della mutazione alla sua rapida affermazione nell'intera popolazione, tra i marcatori si genera un considerevole aumento di linkage disequilibrium. Quando oramai la mutazione è presente in oltre il novanta percento gli effetti della selezione risultano trascurabili e gli eventi di ricombinazione progressivamente rilassano il linkage disequilibrium fino a ripristinare la condizione precedente la comparsa della mutazione.

Il linkage disequilibrium come rilevatore di effetti di selezione genetica risulta quindi molto sensibile alla selezione in corso o recente ma tende anche a "dimenticare" rapidamente una tale pressione.

Diverso è il comportamento di indicatori derivati da un modello a coalescente. Secondo questo modello teorico la storia di un marcatore neutrale in un campione di una popolazione può essere descritta in termini di un solo indicatore, il *rescaled mutation rate*. Avendo prelevato campioni in diverse fasi del processo di selezione, questo indicatore risulta molto sensibile nel rilevare lo stato di selezione del particolare campione. Da sottolineare che la traccia dell'avvenuta selezione non viene persa nei campioni provenienti dalla simulazione e presi quando ormai il linkage disequilibrium non era più informativo.

Nel caso delle popolazioni reali l'analisi di linkage disequilibrium presenta risposte contraddittorie. Sembrerebbero evidenti tracce di selezione solamente nelle razze Marchigiana e Piemontese.

Il dato per la Marchigiana però non è in accordo con le informazioni in letteratura sul grado ancora iniziale di diffusione in questa razza della mutazione. Questo deriva da un errato campionamento eseguito su questa popolazione. La ricostruzione di campioni rappresentativi della popolazione, via *bootstrap* dal campione originale, restituisce valori di linkage disequilibrium compatibili con un processo di selezione ancora non intensivo.

Al contrario per la razza Belgian Blue l'indice di linkage disequilibrium non rileva alcun effetto di selezione, pur essendo documentata in letteratura e pur essendo al limite della fissazione della mutazione in questa razza.

L'analisi con un modello a coalescente aiuta a chiarire la storia di selezione sulla mutazione del gene della miostatina in queste razze. Infatti sia per la razza Belgian Blue che per la razza Piemontese si osserva un considerevole effetto dovuto a selezione sul *rescaled mutation rate*, mentre risulta decisamente trascurabile per il campione originale della razza Marchigiana.

Per queste razze è possibile quindi riconciliare i risultati delle due analisi statistiche in un quadro storico.

Il campione dalla razza Belgian Blue rappresenta una popolazione che ha subito un intenso processo selettivo ma in epoca sufficientemente remota (prima metà del XX secolo) da annullare nelle popolazioni attuali il linkage disequilibrium. La traccia dell'evento di selezione rimane tuttavia nei parametri del modello a coalescenze.

Il campione della razza Piemontese rappresenta invece una popolazione in piena pressione selettiva e sia il linkage disequilibrium che il modello a coalescente ne danno testimonianza.

Il campione della razza Marchigiana proviene da una popolazione dove ancora non si è avuto un processo di intensa selezione sul gene mutato. Gli effetti di linkage disequilibrium derivano piuttosto dall'errato campionamento. Molto interessante il comportamento del *rescaled mutation rate* che in questo caso risulta scarsamente sensibile al campionamento non attribuendo, correttamente, a questa razza effetti di selezione.

La comparazione dei risultati ottenuti dalle analisi di linkage disequilibrium e di coalescenza nel caso di popolazioni soggette a selezione ha portato quindi a due conclusioni principali.

Innanzi tutto la disomogeneità statistica degli aplotipi osservata nell'analisi di linkage disequilibrium decade molto più rapidamente di quanto non accada per l'impronta lasciata da un processo di selezione sulle frequenze dei singoli alleli. Mentre un indice di linkage disequilibrium "dimentica" rapidamente la storia di selezione a causa degli eventi di ricombinazione del DNA, l'analisi di coalescenza permette di rilevare dinamiche più antiche.

Inoltre l'uso di modelli a coalescenza sembra fornire risultati meno sensibili ai problemi di rappresentatività dei campioni di popolazione e alla loro numerosità. Contrariamente all'analisi di linkage per la quale risulta della massima importanza l'affidabilità del campione.

In ogni caso, comunque, l'utilizzo congiunto delle diverse analisi può contribuire a chiarire il quadro complessivo e risolvere le apparenti contraddizioni, dal momento che ciascuna tecnica racconta un aspetto diverso della dinamica storica delle popolazioni studiate.