Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
BIOLOGIA EVOLUZIONISTICA ED AMBIENTALE

CICLO XXIV

COORDINATORE Prof. Guido Barbujani

LE SIMULAZIONI DEL PROCESSO COALESCENTE IN GENETICA DI POPOLAZIONI: INFERENZE DEMOGRAFICHE ED EVOLUTIVE

Settore Scientifico Disciplinare BIO/18

Dottorando
Dott. Benazzo Andrea

_____
(firma)

Tutore
Prof. Bertorelle Giorgio

_____
(firma)

Anni 2009/2011

INTRODUZIONE

1. I modelli matematici e la simulazione del polimorfismo a livello del DNA

La genetica di popolazioni è una disciplina il cui scopo primario è lo studio dei fenomeni che plasmano e influenzano la variabilità genetica nelle popolazioni. Conoscere i meccanismi che da un lato portano alla formazione di nuove varianti alleliche e dall'altro ne comportano la scomparsa, è di fondamentale importanza per la formulazione di modelli matematici che li possano descrivere. I vantaggi di avere modelli che approssimano la realtà sono molteplici: permettono di studiare come queste forze interagiscano tra loro, come abbiano agito nel passato e inoltre di prevedere che effetti avranno, sulla variabilità osservabile a livello del DNA, nel futuro.

Un esempio è lo studio del processo mutazionale. Il cambiamento delle basi azotate in alcuni punti del DNA, è la forza che provoca la comparsa di nuovi alleli in una popolazione. Questi cambiamenti possono avvenire nelle cellule somatiche, e perciò influenzare il comportamento delle cellule nei tessuti differenziati, oppure avvenire nelle cellule germinali ed essere trasmessi alla progenie. Solo questi ultimi sono di interesse evolutivo perché la loro storia può essere studiata nel tempo. La presenza simultanea nella popolazione di due o più alleli ad un locus genomico è detta polimorfismo e generalmente viene generato dalla comparsa di mutazioni in singole posizioni del genoma, denominate "*Single Nucleotide Polymorphism*" (SNP).

La misura più semplice per quantificare la distanza tra due sequenze diverse di DNA, provenienti da un particolare locus genomico, è il numero di differenze nucleotidiche osservabili. Inoltre, se si assume che le mutazioni compaiano ad una frequenza costante nel tempo (orologio molecolare stretto), il numero di differenze è direttamente proporzionale al tempo intercorso dalla loro divergenza. In questo modo però non è tenuta in considerazione la possibilità che un polimorfismo possa essere originato dall'accumulo di più di una mutazione nel tempo, o che un sito uguale tra le due sequenze sia in realtà il risultato di una retromutazione o di una mutazione parallela (vedi Figura1).

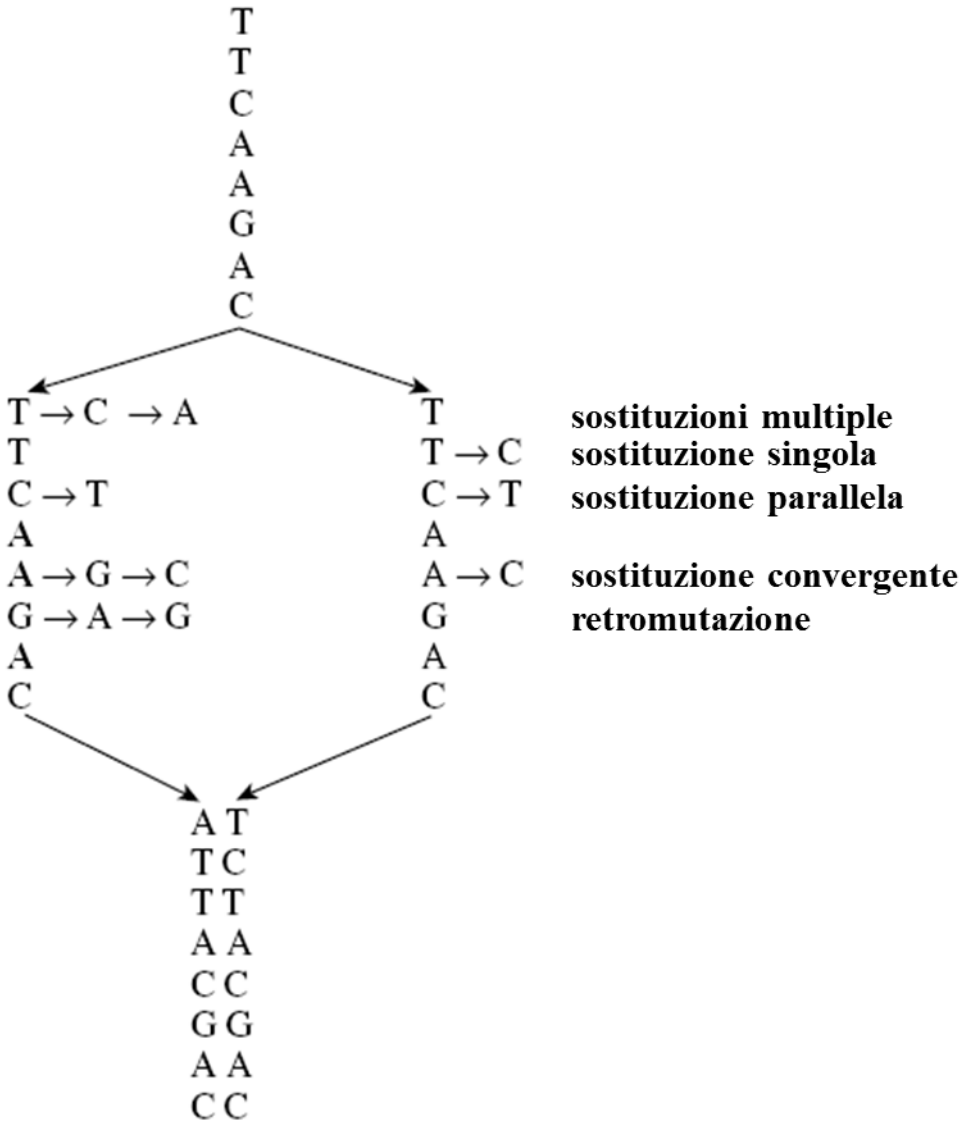

Figura1: Illustrazione dei vari tipi di sostituzioni che possono avvenire in una sequenza di DNA. Una sequenza ancestrale (in alto) diverge in due sequenze che accumulano sostituzioni in modo indipendente (in basso).

Nel 1969, Jukes e Cantor (Jukes e Cantor 1969) proposero un modello per descrivere probabilisticamente un meccanismo di sostituzione nucleotidica che tenesse in considerazione tutti questi elementi. Secondo il loro modello (denominato JC69), ogni sito che compone una sequenza di DNA evolve in maniera indipendente e il cambiamento nucleotidico in ogni posizione nel tempo è descritto da una catena di Markov in tempo continuo, dove le quattro basi azotate del DNA sono i quattro possibili stati della catena e ogni nucleotide ha la stessa probabilità di mutare in uno qualsiasi degli altri tre (Figura2).

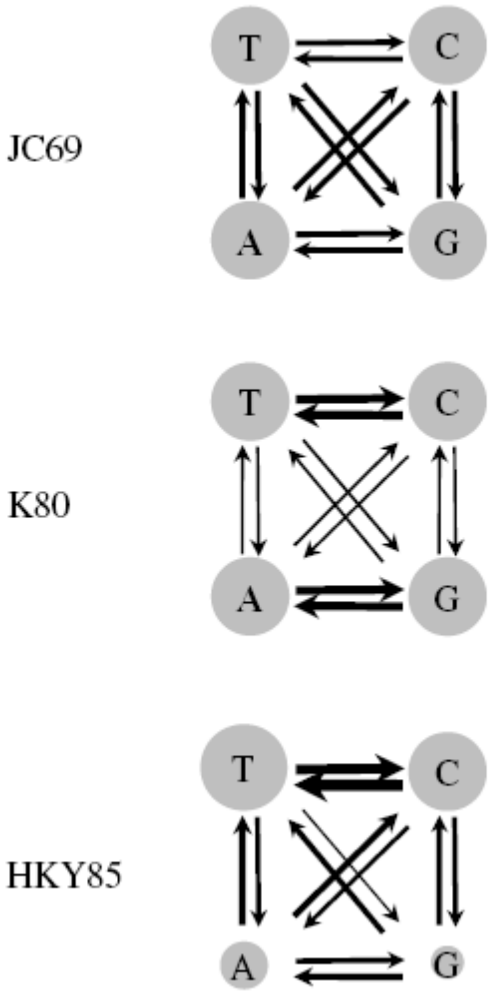

Figura2: Tassi relativi di sostituzione nucleotidica secondo tre diversi modelli mutazionali: JC69 (Jukes e Cantor 1969), K80 (Kimura 1980) e HKY85 (Hasegawa et al. 1985). Lo spessore delle frecce è proporzionale al tasso di sostituzione mentre l'ampiezza dei cerchi rappresenta la frequenza dei nucleotidi nelle distribuzioni all'equilibrio nel modello di Markov.

Utilizzando questa formulazione, è possibile calcolare la probabilità di osservare ognuna delle quattro basi azotate a t = n dato un nucleotide di partenza a t = 0, ed estendendo questo meccanismo a tutti i siti di una sequenza di DNA è possibile prevedere come questa si evolva nel tempo.

Nel corso degli anni sono state sviluppate numerose evoluzioni di questo primo modello di sostituzione (vedi Kimura 1980; Hasegawa et al. 1985; Tamura e Nei 1993) aggiungendo di volta in volta nuove caratteristiche per rendere il modello più realistico, fino ad arrivare alla definizione del "*General Time Reversible*" (Tavaré 1986; Yang 1994; Zharkikh 1994), un modello generale che comprende al suo interno tutti gli altri modelli come casi speciali.

La definizione di questi modelli ha, di fatto, aperto le porte all'uso delle simulazioni in

genetica di popolazione. Infatti, da questo momento è stato possibile "generare" al calcolatore sequenze di DNA evolute secondo un processo che replica quello che succede in natura, e ottenere precise previsioni sul livello di polimorfismo secondo diversi parametri del modello.

In generale è possibile definire una *simulazione al computer* (chiamata anche *simulazione stocastica* o *simulazione Monte Carlo*) come un esperimento virtuale, dove si cerca di riprodurre un processo biologico in un calcolatore per studiarne le proprietà. In alcuni casi viene utilizzato il termine specifico "*simulazione Monte Carlo"* quando il risultato ottenuto è deterministico, come ad esempio nel calcolo integrale, o il termine "*simulazione stocastica"* quando il risultato contiene variazioni casuali. Queste distinzioni non saranno prese in considerazione in questa tesi.

2. I modelli demografici "*forward in time*"

I modelli mutazionali descrivono come le mutazioni insorgono nei singoli individui, ma non tengono in considerazione come la demografia possa influenzarne la diffusione nelle popolazioni. Ad esempio, nelle ricostruzioni filogenetiche si è interessati a studiare eventi di divergenza tra specie avvenuti centinaia di migliaia, o milioni, di anni nel passato tramite l'analisi di porzioni molto conservate del genoma. In questo caso, è ragionevole pensare che eventi demografici di relativa breve durata, come una riduzione o un'espansione demografica, non abbiano lasciato segni significativi a livello del DNA e possano perciò essere ignorati. Quando si analizzano invece gli individui appartenenti ad una popolazione, il livello di polimorfismo è altamente influenzato dalla demografia passata, ed perciò di fondamentale importanza tenere in considerazione la variazione delle caratteristiche della popolazione nel tempo. I modelli demografici rispondono a questa esigenza, descrivendo in maniera formale come gli individui di una popolazione, generalmente rappresentati da un numero N di cromosomi o singoli loci di DNA, evolvono nel tempo.

Nella categoria di modelli *forward in time*, la composizione della popolazione viene modificata generazione dopo generazione secondo le regole del modello, con uno scorrere del tempo dal presente verso il futuro.

Il modello neutrale di Wright-Fisher (Fisher 1930; Wright 1931) è il modello più semplice per descrivere la dinamica di una popolazione costante nel tempo. Il modello assume che la popolazione sia di dimensioni finite, con un numero di individui pari ad N se si considerano organismi aploidi o $2N$ se diploidi. Gli incroci devono essere casuali (popolazione panmittica) e ogni individuo deve avere la stessa probabilità di lasciare discendenti (assenza di selezione). Infine il tempo viene misurato in generazioni discrete e non sovrapposte, cioè ogni generazione è composta solo dai discendenti della generazione precedente. Se queste assunzioni sono rispettate, la composizione della popolazione alla generazione t+1 (futuro) rappresenta un campionamento con

rimpiazzo degli individui dalla generazione t (presente), come illustrato in Figura3.

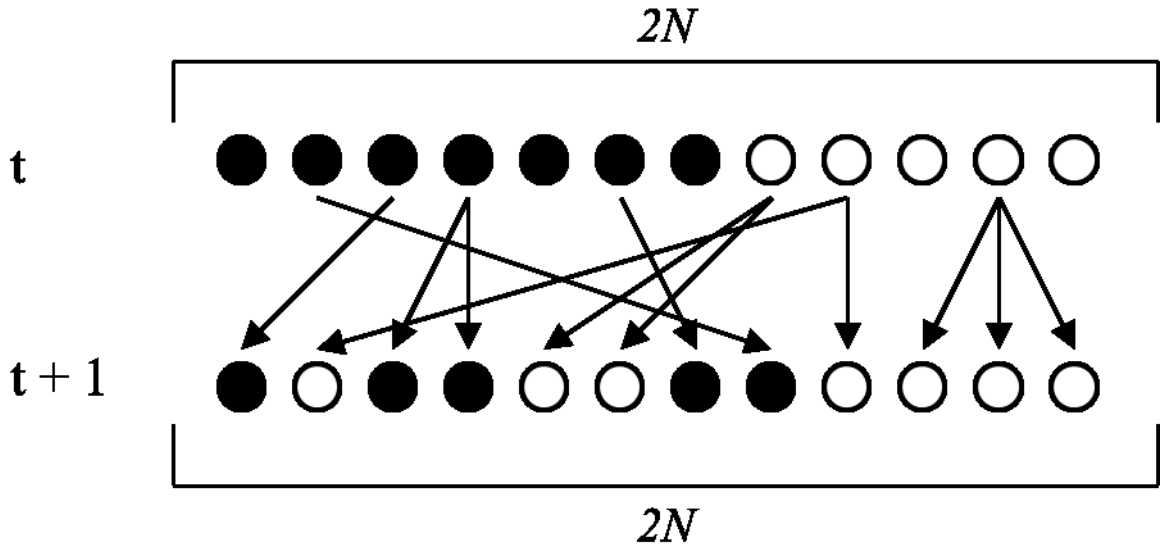

Figura 3: Evoluzione di una popolazione di Wright-Fisher. Le frecce indicano la transizione dalla generazione t alla generazione t + 1, pescando con rimpiazzo *2N* individui dalla generazione t.

Seguendo questo modello, possiamo ad esempio studiare la dinamica delle frequenze alleliche nel tempo. Questo fenomeno, chiamato *deriva genetica*, non è altro che il cambiamento della frequenza di un allele nella popolazione dovuto al campionamento casuale degli individui che si riproducono e che lasciano discendenti nelle generazioni successive. La probabilità di un allele di fissarsi (tutti gli individui della popolazione possiedono quell'allele) o di scomparire è direttamente proporzionale alla dimensione della popolazione. Basandosi su questo modello, nel 1931 Wright introdusse il concetto di *dimensione effettiva* di una popolazione, che permise di confrontare l'impatto della deriva genetica in popolazioni diverse e venne definita come la dimensione di una ipotetica popolazione di Wright-Fisher che sperimenti lo stesso livello di deriva genetica della popolazione studiata. Questo parametro, ad oggi, rimane una delle misure cardine in genetica di popolazioni.

Crow e Kimura (1970) utilizzarono questo modello per studiare, tramite simulazioni, come il tempo necessario per la fissazione di un allele nella popolazione sia direttamente proporzionale alla sua dimensione effettiva, con un tempo atteso di fissazione minore in popolazioni di piccole dimensioni.

Sebbene il modello di Wright-Fisher descriva bene diversi aspetti di una popolazione, contiene alcune assunzioni che ne limitano le capacità di descrivere la dinamica di una popolazione reale. Ad esempio in molte specie la generazione parentale e quella dei discendenti è sovrapposta, per cui in un dato momento temporale la popolazione è formata da una percentuale variabile delle due parti. Inoltre in molti casi reali le popolazioni non mantengono una dimensione costante nel

tempo ma possono andare incontro ad espansioni o contrazioni demografiche. Non tenere in considerazione il fenomeno della selezione naturale è un grosso limite del modello e ne pregiudica l'applicabilità a tutti quei casi in cui questa agisce.

Con il progressivo aumento dei marcatori genetici a disposizione, seguito dalla maggior capacità di calcolo dei computer, si è arrivati alla possibilità di simulare singoli marcatori o l'intero genoma di ogni individuo secondo scenari evolutivi realistici (Hoban et al. 2011). Questo tipo di simulazioni permette di definire lo stato iniziale della popolazione (t=0) assegnandole un certo numero di individui oppure suddividendoli in sottopopolazioni per ricreare una popolazione strutturata al suo interno. Il passaggio alle generazioni successive (t+1, t+2, ecc...) avviene tramite l'accoppiamento degli individui che può essere definito come casuale o non casuale. A questo livello si possono introdurre caratteristiche come l'accoppiamento assortativo, le generazioni sovrapposte, la selezione o la varianza del successo riproduttivo, permettendo ad alcuni individui di lasciare un numero maggiore (o minore) di discendenti nella generazione successiva.

Potenzialmente, l'intero genoma di ogni individuo può essere simulato, perciò è possibile ricreare realistici meccanismi mutazionali e di ricombinazione e assegnarli in modo differenziale lungo il genoma.

3. I modelli demografici "*backward in time*"

Intuitivamente l'evoluzione è associata allo scorrere del tempo. Circa 4 miliardi di anni fa le prime molecole hanno iniziato ad evolversi in strutture sempre più complesse e a conquistare praticamente qualsiasi habitat disponibile sulla Terra, cambiando in continuazione ed accumulando diversità genetica.

Il modello di Wright-Fisher, descritto in precedenza, cerca di riprodurre questo tipo di evoluzione in una scala più piccola, descrivendo le dinamiche dei geni nel tempo, passando da una generazione alla successiva. In questo modo è possibile seguire la linea di discendenza di un gene in un punto qualsiasi del tempo. Ricostruire cosa è successo ad un gene "all'indietro nel tempo" è l'idea dei modelli *backward in time,* che hanno trovato la loro più importante descrizione formale matematica nel modello Coalescente. Su questo modello si basano la maggior parte dei metodi moderni di indagine delle dinamiche di popolazione basate sul DNA (Rosenberg e Nordborg 2002).

Introdotto da Kingman (Kingman 1982a; Kingman 1982b), il Coalescente è un modello stocastico che descrive i rapporti di discendenza dei geni provenienti da un campione della popolazione procedendo dal presente al passato. Dato un campione di N geni ($2N$ nel caso di organismi diploidi), e assumendo che la popolazione si sia mantenuta costante nel tempo, è possibile costruire una genealogia che ne descrive le relazioni di discendenza tracciando quando

avviene un evento di coalescenza, cioè quando due geni hanno un antenato comune nella generazione precedente. La storia di un campione di dimensioni *N* comprende quindi esattamente *N*-1 eventi di coalescenza, alla fine dei quali si avrà un unico antenato comune di tutto il campione, definito come antenato comune più recente (TMRCA). Il risultato finale è un albero come quello mostrato in Figura4.

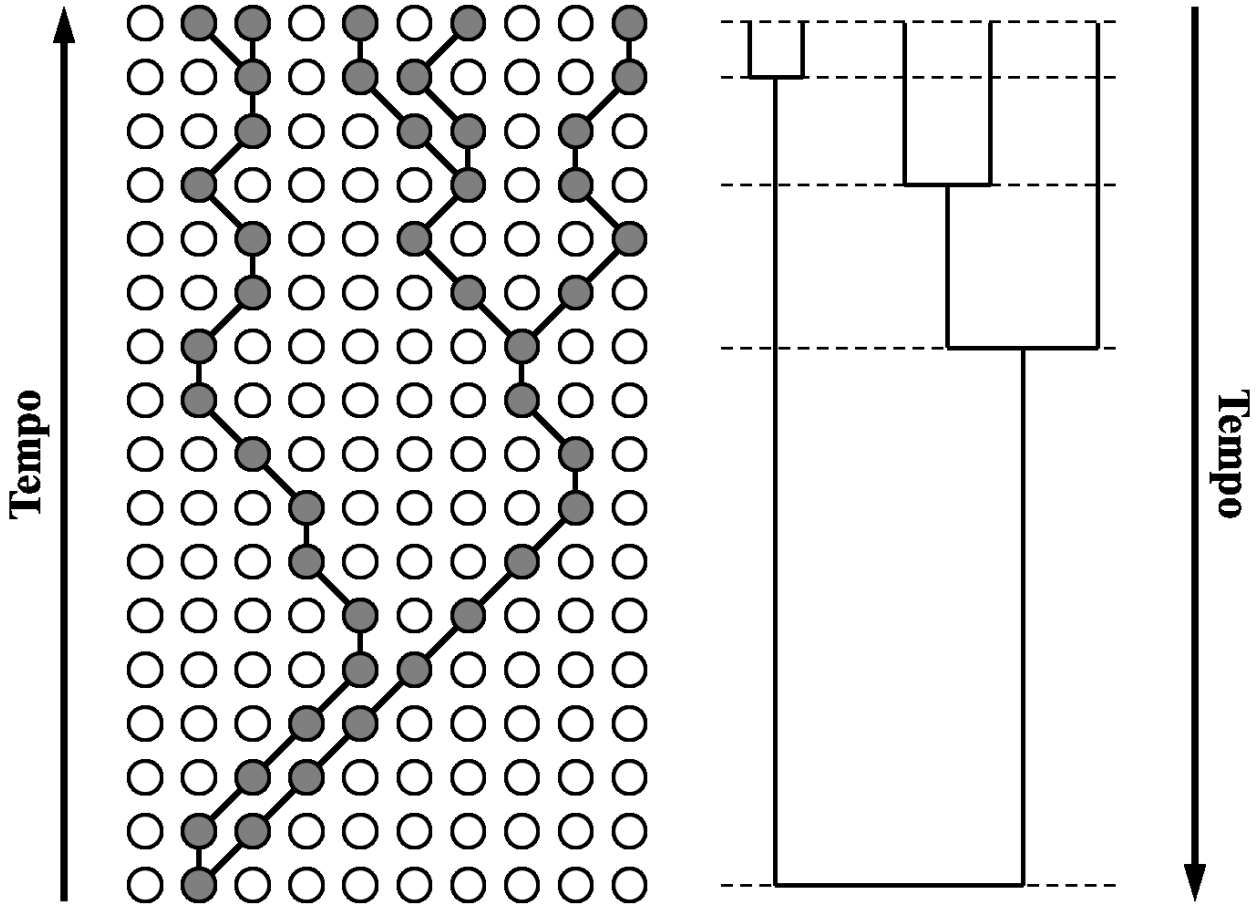

Figura4: A sinistra, una genealogia di cinque sequenze campionate da una popolazione di 10 individui in una popolazione di Wright-Fisher di dimensioni costanti. A un certo tempo, le sequenze si fondono (coalescono) in un unico antenato comune (TMRCA). A destra, è riportato un esempio della stessa genealogia prodotta secondo il modello Coalescente. Il tempo nel modello di Wright-Fisher è discreto mentre nel Coalescente è misurato in modo continuo.

Ogni genealogia, o realizzazione del processo coalescente, viene perciò descritta dalla combinazione di relazioni di parentela tra i membri del campione e dai tempi di coalescenza. La probabilità che due geni abbiano un antenato comune alla generazione precedente è inversamente proporzionale alla dimensione della popolazione da cui provengono. Per questo motivo i cambiamenti demografici influenzano i tempi di coalescenza.

Se consideriamo ad esempio un campione di sequenze di DNA proveniente da una

popolazione che ha subito un'espansione demografica, il tempo atteso prima del primo evento di coalescenza dipenderà dalla dimensione della popolazione moderna e perciò sarà grande. Al contrario, procedendo indietro nel tempo, la dimensione della popolazione sarà più piccola, e così due sequenze potranno avere una coalescenza in tempi più rapidi. L'effetto di questo processo sulla genealogia finale sarà di produrre un albero con rami terminali lunghi e corti rami interni se confrontata con quella attesa secondo uno scenario di popolazione costante (vedi Figura5).

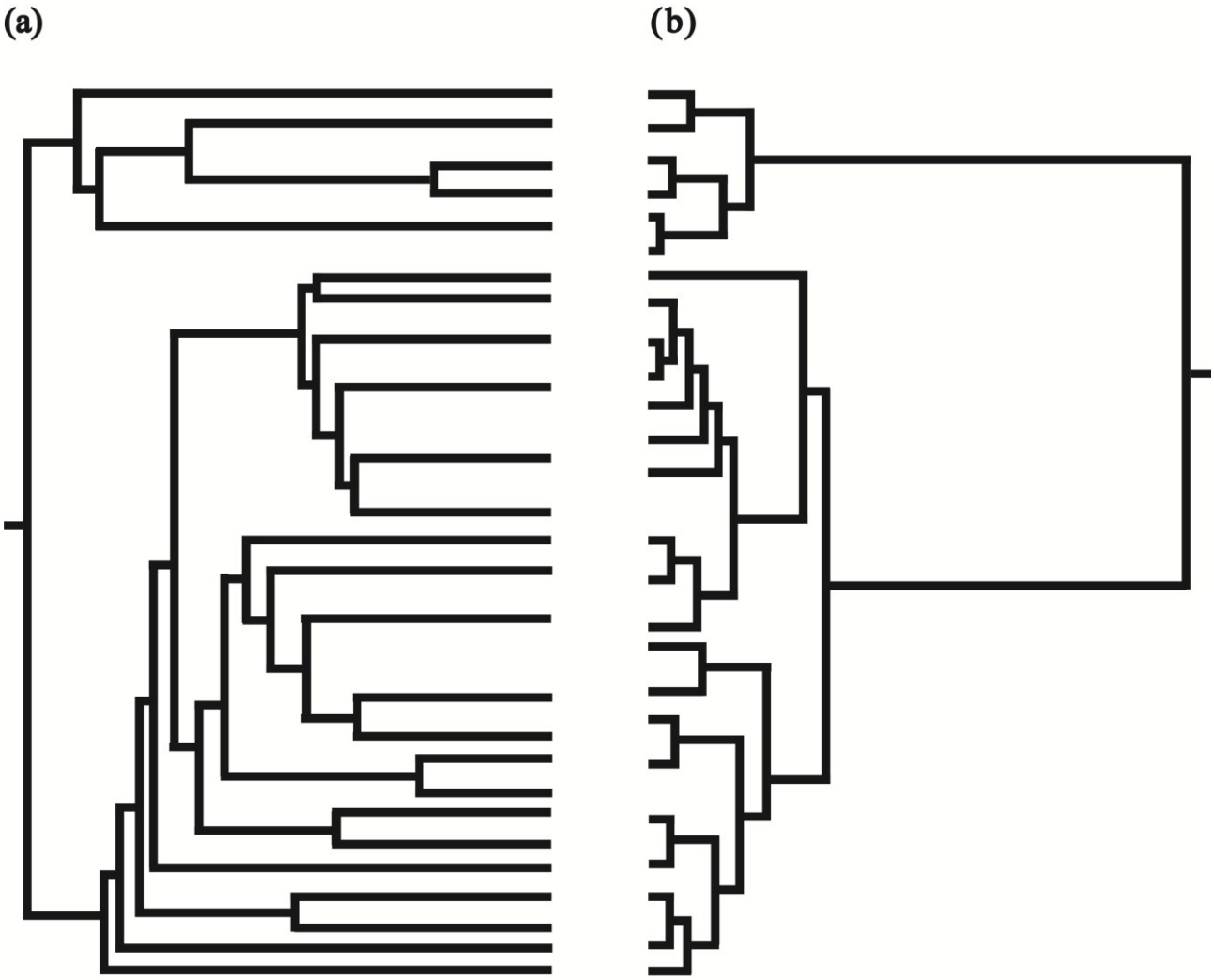

Figura5: Due esempi di genealogie. (a) una genealogia di 25 sequenze campionate da una popolazione che ha subito un espansione demografica; (b) una genealogia con lo stesso numero di sequenze, campionate da una popolazione costante.

Allo stesso modo una popolazione che ha subito un declino demografico sarà caratterizzata da un'elevata frequenza di eventi di coalescenza in tempi recenti dovuta alla dimensione di popolazione moderna piccola per poi avvenire a più bassa frequenza quando la popolazione diventa

più grande nel passato. Utilizzando questo modello è quindi possibile simulare in maniera semplice e veloce una genealogia di un locus genomico di interesse secondo uno scenario demografico predefinito. Una volta ricostruita la genealogia è possibile simulare la variabilità genetica a quel locus utilizzando un qualsiasi modello mutazionale inserendo le mutazioni in maniera proporzionale alla lunghezza dei rami.

Negli anni successivi alla sua introduzione, il Coalescente è stato esteso in modo da includere fenomeni popolazionistici e genetici che ne hanno di fatto aumentato la capacità di generare variabilità genetica secondo scenari demografici realistici.

Ad oggi esistono decine di simulatori con diverse caratteristi e che danno la possibilità di simulare popolazioni multiple, cambiamenti demografici, modelli spaziali espliciti, migrazione fra le popolazioni, ricombinazione, diversi modelli mutazionali, selezione e il campionamento antico (vedi review di Hoban et al. 2011 per una descrizione dettagliata).

4. Le simulazioni in genetica di popolazioni

In genetica di popolazioni si è spesso interessati allo studio di sistemi (individui, popolazioni o specie) sui quali agiscono molteplici fenomeni naturali o artificiali (ad esempio indotti dall'uomo), ed è spesso difficile capire come questi interagiscano fra loro.

Le simulazioni al calcolatore sono ottimi strumenti per capire il funzionamento di questi sistemi complessi. In principio generale è quello di generare dataset *in-silico* di variabilità genetica secondo specifiche combinazioni di scenari evolutivi e caratteristiche genetiche. Ad esempio, se applicato allo studio dell'evoluzione umana (vedi Ray e Excoffier 2009), lo scenario demografico può contenere espansioni demografiche da particolari regioni geografiche, differenti dimensioni di popolazione antiche e recenti e livelli di migrazione entro e tra continenti. Inoltre, nella simulazione possono essere specificati diversi tipi di marcatori genetici, tassi di mutazione e assunzioni sull'associazione di diversi marcatori.

Il processo di simulazione normalmente prevede la ripetizione di una simulazione per tenere in considerazione la stocasticità del processo demografico, mutazionale e del campionamento. I dataset prodotti in questo modo di solito sono utilizzati in tre classi di applicazioni in genetica di popolazioni: scopi predittivi, l'inferenza statistica e la validazione di metodi statistici.

Scopi predittivi

Le simulazioni per definizione predicono lo stato futuro di un sistema secondo in condizioni specifiche predeterminate e perciò sono diventate uno strumento molto affermato soprattutto in biologia evoluzionistica per studiare problemi complessi come l'evoluzione della riproduzione

sessuale o la teoria della speciazione (Lively et al 2010; Church et al 2002).

Oltre che in questo campo, le simulazioni sono molto utilizzate anche in genetica della conservazione e in epidemiologia. Ad esempio, la conservazione, gestione e recupero di specie in pericolo di estinzione (o importanti dal punto di vista economico) è spesso complicata dall'influenza di molti fattori tra cui i cambiamenti climatici, le malattie e lo sovra sfruttamento. In questo ambito, l'obiettivo principale è quello di prevedere l'impatto di cambiamenti ecologici (frammentazione dell'habitat, epidemie di malattie, ecc..) o dell'intervento umano (traslocazioni di individui, allevamento, ecc..) sulla variabilità genetica futura. L'approccio generalmente seguito è quello di simulare una popolazione secondo diversi scenari di flusso genico, controllare il livello di variabilità genetica prodotto e infine identificare i valori dei parametri che permettono di raggiungere l'obiettivo a cui si è interessati. Ad esempio, un problema molto frequente è determinare quale tasso di migrazione garantisca il livello desiderato di eterozigosità, differenziamento genetico o outbreeding (Daleszczyk 2009; Vonholdt 2008) in una popolazione. In altri casi questo tipo di approccio è stato utilizzato per valutare le conseguenze genetiche del mantenere una barriera tra le popolazione di bisonte Europeo rispetto alla condizione di assenza di barriera (Daleszczyk 2009); determinare la minima dimensione effettiva di popolazione, e tassi di migrazione, per mantenere livelli accettabili di eterozigosità nel lupo di Yellowstone (Vonholdt 2008); cercare il numero di colonie necessarie per mantenere variabilità genetica negli alleli sessuali in popolazioni di api controllate (Alves et al. 2011); e predirre l'impatto genetico di riduzioni demografiche assumendo diverse intensità e tempi di reintroduzione di individui, caccia di frodo o eventi naturali (Kenney et al. 1995; Ng et al. 2009).

L'utilizzo delle simulazioni si è dimostrato altrettanto importante in genetica medica. In questo campo esiste un forte interesse nel riuscire ad identificare loci del genoma associati alla suscettibilità alle malattie. Una teoria sull'architettura genetica delle malattie prevede che le malattie comuni siano causate da pochi alleli condivisi tra le popolazioni ("*Common Disease Common Variants*", CDCV), ma le condizioni genetiche e demografiche necessarie perché questo avvenga sono molto discusse. Simulazioni in forward sono state usate per testare questa ipotesi secondo uno scenario di popolazione stabile dal punto di vista demografico, o in espansione, con vari livelli di mutazione, migrazione e selezione (Peng e Kimmel 2007).

Le simulazioni hanno mostrato come un modello definito utilizzando le informazioni disponibili sulla storia evolutiva umana (una recente espansione demografica) riesca a spiegare gli alti livelli di diversità genetica delle malattie rare e perciò supporti l'ipotesi CDCV.

Inferenza statistica

Negli ultimi decenni si è verificato un progressivo interesse nel far luce sulla storia evolutiva delle specie, ed in particolare, nell'identificare cambiamenti demografici, il mescolamento e la divergenza tra le popolazioni e nelle modifiche degli areali di distribuzione indotti dai cambiamenti climatici (Fagundes et al. 2007; Ray e Excoffier 2009). Le simulazioni in questi anni sono diventate uno strumento essenziale per lo studio di questi fenomeni. Un problema comune in questo campo è confrontare ipotesi demografiche alternative e di stimare parametri demografici e genetici alla base dei modelli maggiormente supportati dai dati. Un approccio molto diffuso è quello di generare dati genetici secondo diverse storie evolutive, utilizzare alcune statistiche descrittive (come l'Fst, numero di alleli, ecc..) per riassumere le informazioni sul polimorfismo e creare, infine, le distribuzioni degli indici calcolati generate secondo ogni modello. Lo scenario migliore viene successivamente identificato confrontando le distribuzioni degli indici calcolati nelle simulazioni e nei dati reali. Questo ultimo passaggio viene eseguito utilizzando diversi metodi sviluppati *ad-hoc*, o più recentemente tramite l'Approximated Bayesian Computation, che ha di fatto rivoluzionato l'uso delle simulazioni per l'inferenza statistica (Beaumont et al. 2002; Beaumont 2010; Csillery et al. 2010).

Numerosi esempi dell'utilizzo delle simulazioni per l'inferenza statistica sono presenti in letteratura e spaziano dalla biologia evolutiva, all'ecologia, alla conservazione e all'epidemiologia. In conservazione, ad esempio, questo tipo di approccio è stato applicato allo studio dell'invasione di un nuovo habitat da parte di una specie. Gli eventi di colonizzazione sono importanti fattori di cambiamento ecologico ed evoluzionistico ed è perciò di particolare interessa studiare come il fenomeno della deriva genetica, e della selezione naturale, agiscano a livello del DNA e influenzino il successo di un invasione. La complessità di questi eventi, però, pone dei seri limiti agli approcci tradizionali basati sulla massima verosimiglianza, o Bayesiani, ma non quelli basati sulle simulazioni. Alcuni esempi includono la stima del numero e dell'origine degli individui fondatori di rana toro che hanno dato origine alla colonizzazione del Nord America (Ficetola et al. 2008) oppure lo studio del flusso genico tra alcune popolazioni di ricci di mare recentemente formatesi e l'ipotetica popolazione sorgente in Australia (Banks et al. 2010).

Le simulazioni possono essere uno strumento utile anche nella fase di pianificazione di un esperimento. Lo schema di campionamento (cioè il numero di marcatori e il numero di campioni) adeguato per identificare un fenomeno attraverso un test statistico è uno dei problemi più comuni in genetica di popolazioni: un campione troppo piccolo potrebbe non contenere abbastanza informazione per identificare il fenomeno che stiamo cercando, mentre produrre un campione troppo grande significa spendere più risorse del necessario.

A questo scopo si possono effettuare simulazioni di dati genetici secondo diversi schemi di

campionamento alla ricerca della miglior combinazione che permetta di identificare un particolare fenomeno con una certa confidenza. Ad esempio, con questo approccio è stato verificato che l'accuratezza nella stima di massima verosimiglianza di θ aumenta più velocemente incrementando il numero di loci inclusi nell'analisi rispetto all'incremento degli individui campionati o dell'aumento della lunghezza delle sequenze di DNA analizzate (Felsenstein 2006).

Validazione di metodi statistici

Di pari passo con l'aumento della quantità di dati genetici disponibili, vengono sviluppati nuovi metodi e strumenti statistici per estrarre informazioni dai dati di variabilità genetica. Una volta proposto un nuovo metodo o strumento, le simulazioni sono generalmente utilizzate per verificare come questo si comporta in particolari condizioni.

Un aspetto molto importante è verificare l'efficienza di un metodo nel raggiungere il suo scopo. L'approccio generalmente utilizzato prevede di definire un intervallo prefissato di valori per un parametro d'interesse. In seguito, una o più simulazioni vengono effettuate secondo quel parametro, i dati prodotti vengono analizzati con la metodologia proposta e, infine, il potere del metodo viene calcolato insieme ad alcune misure di qualità della stima. Questo approccio è stato seguito ad esempio per verificare il potere di un metodo nell'identificare una riduzione della dimensione effettiva di una popolazione (Luikart et al. 1998) o il potere di alcuni test di assegnazione secondo diversi livelli di struttura di popolazione e variabilità genetica (Manel et al. 2002).

Lo stesso approccio può essere utilizzato per confrontare l'abilità di diversi metodi nel raggiungere lo stesso obbiettivo: ad esempio è stato utilizzato per verificare come un semplice nuovo metodo basato sull'analisi delle componenti principali (DAPC, Jombart et al. 2010) abbia una capacità maggiore nell' identificare la struttura nelle popolazioni rispetto a un più complesso metodo bayesiano (Pritchard et al. 2000).

In molti casi, un metodo (o un modello) necessita che alcune assunzioni siano rispettate per poter essere applicato ed ottenere una stima credibile. Non rispettare le assunzioni può influenzare il risultato della metodologia più o meno pesantemente. Le simulazioni possono essere usate per verificare quanto un metodo sia "robusto" quando le sue assunzioni vengono violate. Per far questo, vari livelli di intensità di un fenomeno vengono simulati e sui dati prodotti dalle simulazioni viene applicato il metodo. Più si osserva uno scostamento dalla verità, più il metodo è influenzato da quel particolare fenomeno e, in questo modo, è anche possibile classificare quei fenomeni che lo influenzano maggiormente. Questo approccio è stato utilizzato ad esempio per verificare come la stima di parametri demografici secondo il modello "Isolation with Migration" (Nielsen e Wakeley

2001) si dimostri robusta alla violazione di numerose assunzioni come l'assenza di migrazione, la presenza di ricombinazione intra-locus o migrazione da popolazioni non campionate, mentre risulta essere molto sensibile a modificazioni del modello mutazionale (Strasburg e Rieseberg 2010).

5. Scopo della tesi

In questo studio mi sono occupato dell'utilizzo delle simulazioni genetiche, basate sul Coalescente, applicate a quattro diversi problemi di genetica di popolazioni che rientrano nei campi applicativi descritti nell'introduzione.

Nella prima applicazione, ho effettuato uno studio di simulazioni applicato ad una metodologia statistica, chiamata "Bayesian Skyline Plot" (Drummond et al. 2005), che permette di ricostruire la dinamica della dimensione effettiva di una popolazione nel tempo. Un'assunzione del modello alla base di questa metodologia prevede che la popolazione in esame sia isolata, cioè non invii o riceva migranti da un'altra popolazione. Con un approccio simile a quanto descritto nella sezione 1.4.3 ho studiato gli effetti della violazione dell'assunzione d'isolamento sulla ricostruzione della dimensione di popolazione nel tempo.

Nel secondo studio ho utilizzato le simulazioni per quantificare l'importanza dell'inclusione di campioni antichi all'interno di un'analisi demografica mirata ad identificare una riduzione demografica e di identificare gli schemi di campionamento (combinazione di DNA antico e moderno) ideali per massimizzare la probabilità di identificare il fenomeno demografico minimizzando il costo di produzione del campione.

Nel terzo studio ho collaborato alla stesura di una review sulle applicazioni di un metodologia statistica basata sulle simulazioni, chiamata Approximate Bayesian Computation (Beaumont et al. 2002). La review si divide in due parti principali: nella prima parte sono state descritte le origini della metodologia e come si è sviluppata nel tempo. Nella seconda, sono stati discussi in maniera critica tutti i passaggi necessari alla realizzazione di un'analisi completa e i lavori usciti dal 1997 al 2009 in cui questa metodologia è stata impiegata.

Nel quarto studio ho applicato l'Approximated Bayesian Computation per indagare la storia demografica di tre specie di pesci antartici appartenenti al genere *Chionodraco*. I dati genetici sono stati utilizzati per ricostruire eventi di ibridazione interspecifica e come i cicli glaciali abbiano influito sul processo.

APPLICAZIONE NUMERO UNO

L'EFFETTO DELLA MIGRAZIONE SULLA RICOSTRUZIONE DELLA DIMENSIONE EFFETTIVA DI UNA POPOLAZIONE NEL TEMPO: UNO STUDIO DI SIMULAZIONE

1.1 INTRODUZIONE

La storia demografica di una popolazione modella il pattern di variabilità genetica osservabile nel genoma dei suoi componenti moderni. La ricostruzione di questa dinamica, a partire da un campione di DNA moderno, permette perciò di raccogliere importanti informazioni sui processi evolutivi avvenuti come ad esempio studiare la correlazione tra eventi demografici e paleo climatici (Drummond et al. 2005; Campos et al. 2010), analizzare i fattori che hanno influenzato la dinamica delle popolazioni nel passato (Finlay et al. 2007; Atkinson et al. 2008; Stiller et al. 2010) e tracciare la trasmissione e la diffusione di virus (Kitchen et al. 2008; Magiorkinis et al. 2009).
Diversi metodi sono ad oggi disponibili per stimare la storia demografica di un popolazione utilizzando sequenze di DNA (Hey 2010), molti dei quali assumono che essa sia descritta da un semplice modello parametrico come una crescita esponenziale o logistica. La storia demografica poi può essere inferita valutando quanto i dati genetici supportano un particolare modello piuttosto che un altro oppure stimando direttamente i parametri del modello. Ad esempio la stima del tasso di crescita di un modello ad espansione esponenziale può assumere un valore positivo, negativo o uguale a zero, indicando rispettivamente una popolazione in crescita, decrescita o costante.

Molto spesso però la storia demografica di una popolazione è più complessa di quanto descritto da un semplice modello parametrico perciò è stata sviluppata una classe di modelli non parametrici o semiparametrici, denominati "Skyline Plot", in grado di ricostruire le dinamiche popolazionistiche nel tempo a partire da un campione di sequenze di DNA, o da una genealogia stimata in precedenza, senza la necessità di definire a priori un insieme di possibili modelli demografici (Pybus et al. 2000; Strimmer e Pybus 2001; Drummond et al. 2005; Heled e Drummond 2008; Ho e Shapiro 2011).Tutti questi metodi si basano sulla teoria coalescente (Kingman 1982a; Kingman 1982b) per mettere in relazione la genealogia delle sequenze appartenenti al campione e la storia demografica della popolazione da cui proviene. La maggior parte delle proprietà di questi metodi derivano direttamente dal Coalescente includendo dunque anche le assunzioni e le limitazioni. La ricostruzione della dimensione effettiva nel tempo si può scomporre in due passaggi distinti. Nella prima parte, la genealogia che descrive i rapporti genealogici tra gli individui facenti parte del campione è stimata a partire dai dati genetici mentre

nella seconda parte, la genealogia inferita viene utilizzata per ricostruire la storia demografica della popolazione che dipende essenzialmente da il tempo degli eventi di coalescenza presenti nella genealogia. La flessibilità di questa metodologia, insieme allo sviluppo di un software user-frinedly per l'applicazione pratica (BEAST, Drummond e Rambaut 2007), ha fatto si che lo "Skyline Plot" sia diventato uno degli strumenti più utilizzati per ricostruire la storia demografica delle popolazioni.

Una delle assunzioni principali prevede che il campione studiato provenga da una popolazione panmittica isolata, cioè non scambi (o abbia scambiato) migranti con altre popolazioni. L'effetto di includere nel campione alcuni individui migrati da un'altra popolazione si ripercuote direttamente sui tempi di coalescenza stimati e perciò può provocare una distorsione della ricostruzione demografica difficilmente quantificabile ma proporzionale all'intensità della migrazione e al livello di differenziamento delle popolazioni che scambiano migranti. Questo fenomeno è da tenere in considerazione soprattutto quando si studiano specie animali altamente mobili o quando non ci siano barriere conosciute a limitarne il flusso genico tra le popolazioni.
Sebbene la violazione dell'assunzione di panmissia sia stata presa in considerazione in alcuni studi (Shapiro et al. 2004; Drummond et al. 2005; Atkinson et al. 2008; Atkinson et al. 2009), ad oggi non è presente in letteratura un'indagine rigorosa ed esaustiva sulla robustezza dello "Skyline Plot" quando le sue assunzioni non sono rispettate ed in particolare l'effetto della migrazione.
In questo studio mi sono occupato di analizzare attraverso uno studio di simulazione come vari livelli di migrazione influiscano sulla ricostruzione della dimensione effettiva nel tempo utilizzando l'ultimo modello sviluppato, denominato "Extended Bayesian Skyline Plot"(Heled e Drummond 2008). Questo modello offre notevoli vantaggi rispetto ai precedenti, tra i quali la possibilità di analizzare più di un locus nella stessa analisi, di tenere conto della ploidia dei loci genetici e di stimare il numero di cambiamenti demografici avvenuti nel passato. Varie intensità di migrazione sono state prese in considerazione in modo da comprendere valori realistici osservati in popolazioni reali. Ho analizzato inoltre l'effetto delle dimensioni effettive della popolazione che scambia migranti e della diversità degli individui che vengono scambiati sull'analisi. Lo "Skyline Plot" sembra essere robusto a livelli blandi di migrazione, ma in alcuni casi sono stati verificati effetti distorsivi di uno scenario di popolazione costante nel tempo verso scenari di riduzione o espansione demografica.

1.2 MATERIALI E METODI

I dati genetici simulati

I dataset simulati sono stati creati utilizzando il modello coalescente implementato nel software Simcoal (Excoffier et al. 2000). Ogni dataset è composto da 30 individui aploidi provenienti da una popolazione che può ricevere migranti a varie intensità. In base alla condizione testata, ogni individuo può essere rappresentato da un locus, 5 loci o 10 loci di 10.000 paia di basi ognuno. I dati sono stati generati secondo il modello mutazionale HKY85 (Hasegawa et al. 1985) con un tasso di mutazione di 1.25×10^{-7} mutazioni per sito per generazione, tipico di una regione autosomica non codificante (Fagundes et al. 2007). La frequenza dei quattro nucleotidi (A,C,T,G) è stata impostata come uguale e il tasso di trasversioni/transizioni uguale a uno (una mutazione ha la stessa probabilità di essere una transizione o una trasversione). Il modello mutazionale impiegato è a siti finiti, ma la probabilità che due o più mutazioni avvengano sullo stesso sito in una sequenza di 10.000 paia di basi è estremamente bassa, e perciò trascurabile.

Il modello demografico

Il modello demografico utilizzato per la simulazione prevede la presenza nel passato di una popolazione ancestrale, di dimensione *Na*, che si suddivide istantaneamente in due popolazioni di uguale dimensione (*Ne*) al tempo T_{DIV}. Successivamente, le popolazioni appena divise iniziano a scambiare migranti al tempo T_{MIG} fino al presente, con un'intensità regolata dal tasso di migrazione *m* (vedi Figura1.1).

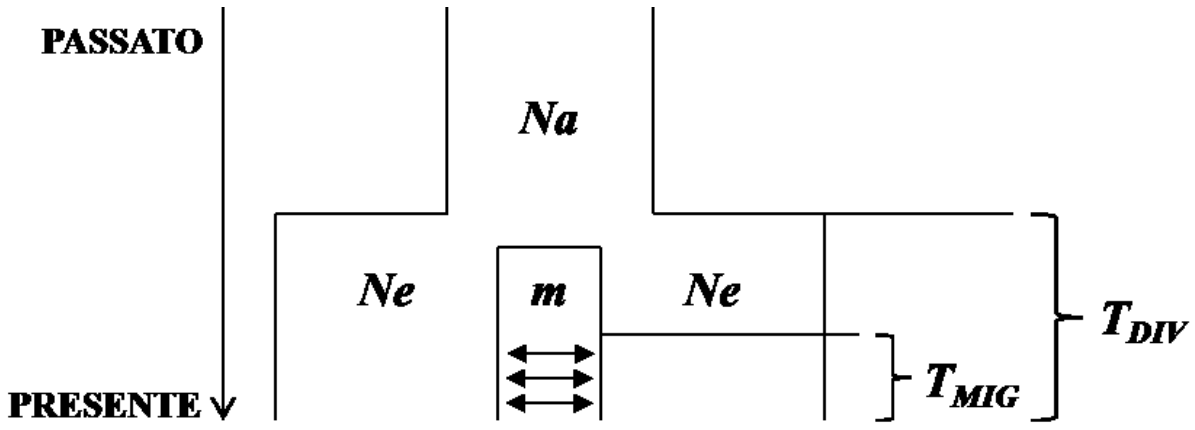

Figura 6.1: il modello demografico utilizzato per le simulazioni coalescenti.

Il modello viene descritto da 5 parametri (*Na*, *Ne*, T_{DIV}, T_{MIG}, *m*), ognuno dei quali è stato variato in modo da esplorare uno spazio formato da 40 combinazioni (vedi Tabella 1.1 e Tabella1.1S per la descrizione delle combinazioni).

Tabella 1.1: I valori dei parametri demografici utilizzati per le simulazioni coalescenti.

Nome	Abbreviazione	Unità di misura	Intervallo valori
Dimensione effettiva ancestrale	*Na*	Individui aploidi	10 000
Dimensione effettiva moderna	*Ne*	Individui aploidi	10 000
Tasso di migrazione	*m*	Tasso per generazione	10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}
Inizio della migrazione	T_{MIG}	Generazioni	50, 150, 2 000, 5 000, 10 000
Tempo di divergenza	T_{DIV}	Generazioni	2 000, 10 000

La dimensione effettiva della popolazione moderna e ancestrale sono state fissate entrambe ad un valore di 10 000 individui aploidi, in modo da mimare una popolazione costante nel corso del tempo in assenza di migrazione. In questo modo, in assenza di migrazione, la storia demografica ricostruita a partire dalle informazioni genetiche degli individui di una sola delle popolazioni moderne dovrà essere costante e proprio questo caso sarà il modello nullo di confronto con scenari dove la migrazione sarà presente. L'intervallo d'intensità di migrazione è stato scelto in modo da esplorare casi dove la migrazione è bassa (10^{-5}x10 000 = 0.1 individui migranti aploidi per generazione) fino a casi dove la migrazione è molto alta (10^{-1}x10 000 = 1 000 individui migranti aploidi per generazione), dove le due sottopopolazioni simulate sono da considerarsi come un'unica popolazione panmittica. Il tempo della divergenza delle due popolazioni moderne è stato scelto in modo da considerare una divergenza recente (T_{DIV} = 2 000) e una divergenza antica (T_{DIV} = 10 000). Allo stesso modo il tempo di inizio della migrazione è stato scelto in modo da esplorare diversi casi: le popolazioni moderne iniziano recentemente a scambiare migranti (T_{MIG} = 50, 150), in tempi intermedi (T_{MIG} = 2 000, 5 000) e in tempi antichi (T_{MIG} = 10 000). Per ognuna delle combinazioni di parametri definite sono state prodotte 100 simulazioni in modo da tenere in considerazione la componente stocastica del Coalescente.

La variabilità genetica simulata

Secondo ogni scenario demografico definito, sono state prodotte 1 000 simulazioni coalescenti per valutare i livelli di variabilità genetica presenti. Il livello di variabilità genetica intra-popolazione è stato valutato attraverso la media, il 2.5° e il 97.5° percentile della distribuzione del numero di siti polimorfici presenti. Il livello di diversità tra le due popolazioni che scambiano migranti è stato analizzato attraverso la media, il 2.5° e il 97.5° percentile della distribuzione di Fst (Cockerham e Weir 1984). Infine sono state calcolate la D di Tajima (Tajima 1989) e la Fs di Fu (Fu 1997) che, in assenza di selezione attiva sulla popolazione, indicano la presenza di cambiamenti demografici passati: il valore 0 è atteso se la popolazione è rimasta di dimensioni costanti nel tempo, valori negativi indicano espansione demografica mentre valori positivi sono indicativi di riduzioni demografiche. Per ognuna delle 1 000 simulazioni è stato verificato se i valori per questi due indici sono significativamente diversi da 0 (bootstrap di 1 000 simulazioni, α=0.05) ed infine è stata riportata la percentuale totale di simulazioni significative per ognuno dei due indici. Le analisi

sopra riportate sono state effettuate con il software Arlequin3.5.1 (Excoffier e Lischer 2010).

La ricostruzione della storia demografica

Tutti i metodi definiti nella classe degli "Skyline Plot" prevedono due passaggi distinti e separabili. La genealogia degli individui analizzati deve essere ricostruita a partire dai dati genetici ed include non solo la stima delle relazioni tra gli individui (topologia dell'albero) ma anche i loro tempi di divergenza (età dei nodi). Questo passaggio può essere effettuato utilizzando i metodi filogenetici standard Bayesiani o di massima-verosimiglianza. Una condizione essenziale è che le lunghezze dei rami dell'albero siano proporzionali con il tempo, perciò il tempo deve essere scalato in mutazioni, anni o generazioni. La genealogia, essendo una stima basata su un campione di sequenze di DNA, porta con se un errore, chiamato "errore filogenetico", che può essere di notevole entità quando la genealogia contiene rami interni corti. Inoltre, molti organismi sono caratterizzati da una bassa variabilità genetica intraspecifica che provoca un aumento della varianza stocastica nella lunghezza dei rami. Nonostante questi fattori, se vogliamo ricostruire la storia demografica di una popolazione, non serve che la genealogia sia ben risolta, soprattutto quando le stime sono pesate tra un grande numero di alberi come nel framework Bayesiano (Drummond et al. 2005).
Il secondo passaggio prevede la stima della storia demografica basata sulla genealogia stimata. Una caratteristica molto utile di questa fase è che dipende solamente dal tempo degli eventi di coalescenza e non dal fatto di ricostruire la genealogia esatta delle sequenze del campione (Pybus et al. 2000). Per esempio, osservare eventi di coalescenza molto vicini tra loro è indicativo di una dimensione effettiva piccola, e questo principio può essere sfruttato per stimare la dinamica della dimensione effettiva. Più precisamente, i metodi "Skyline Plot" si basano sulla semplice relazione tra la dimensione effettiva della popolazione e la lunghezza attesa degli intervalli di coalescenza secondo il modello Coalescente: la dimensione effettiva media in ogni intervallo tra due eventi di coalescenza può essere stimata dal prodotto della lunghezza dell'intervallo (γ_i) e $i(i-1)/2$, dove i rappresenta il numero di linee dell'albero presenti nell'intervallo (Figura1.2a). In questo modo è possibile ottenere una stima della dimensione effettiva in ogni intervallo (definito da ogni evento di coalescenza) della genealogia stimata (Figura1.2b) e così ricostruire la storia demografica tramite la dinamica della dimensione effettiva intervallo dopo intervallo. La ricostruzione demografica include una considerevole parte di incertezza dovuta alla natura stocastica del Coalescente. Infatti, ogni genealogia considerata è solo una singola realizzazione casuale di questo processo e questo comporta che, ad esempio, la stima della dimensione effettiva in ogni intervallo di coalescenza incorpori una notevole quantità di errore. L'errore dovuto al Coalescente è inversamente proporzionale al numero di linee genealogiche presenti in ogni intervallo di coalescenza e quindi

non è uniforme lungo la genealogia: più ci avviciniamo alla radice dell'albero, più aumenta l'errore nella stima della dimensione effettiva. Ad esempio, l'ultimo intervallo di coalescenza (γ_2 in Figura1.2) viene stimato a partire da sole due linee ed è perciò l'intervallo con più errore associato. Questo fatto diventa perciò di notevole importanza ad esempio quando si sta considerando una popolazione costante dove, in media, l'ultimo intervallo di coalescenza occupa metà della genealogia.

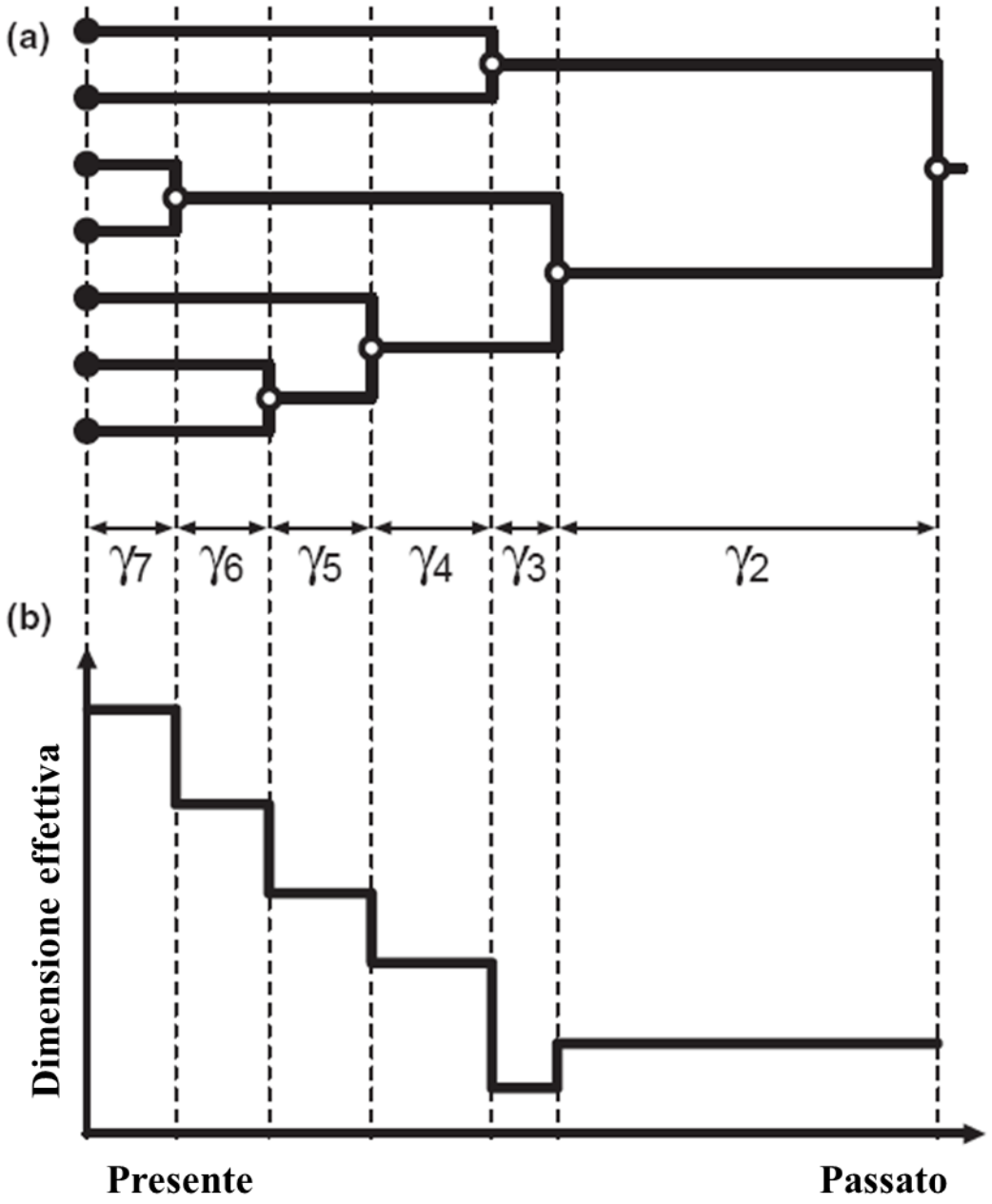

Figura 1.2: Stima della storia demografica da una genealogia. (a) Una genealogia stimata a partire dai dati genetici, dove la lunghezza di ogni ramo è proporzionale al tempo. Gli intervalli di coalescenza sono indicati con Υ_i. (b) La dimensione effettiva della popolazione stimata in ogni intervallo di coalescenza.

Una volta ottenuta la stima della dimensione effettiva in ogni intervallo, la dinamica della

dimensione effettiva nel tempo viene ricostruita in principalmente due modi (vedi review Ho e Shapiro 2011 per i dettagli dei metodi): i) partendo dall'intervallo più recente, la dimensione effettiva viene unita con quella dell'intervallo adiacente assumendo che sia rimasta costante all'interno dell'intervallo, producendo un grafico a "scalini" simile a quello rappresentato in Figura1.2b; ii) la dimensione effettiva non rimane costante all'interno dell'intervallo, ma varia in maniera lineare tra due intervalli, rappresentando in maniera più realistica come una popolazione aumenta o si riduce rispetto a un cambiamento istantaneo. Nei metodi "Skyline Plot" bayesiani la dimensione effettiva in ogni intervallo non è descritta da un singolo valore ma bensì da una distribuzione, detta distribuzione a posteriori. Perciò, per ogni intervallo di coalescenza viene utilizzato un indice di tendenza centrale come la media, moda o mediana come stima della dimensione effettiva. Insieme alla stima, viene riportato anche l' "*High Posterior Density Interval*" (HPD, o intervallo di credibilità) che descrive qual è l'intervallo più piccolo che contiene il 90% o il 95% dei valori a maggior frequenza (vedi Figura1.3).

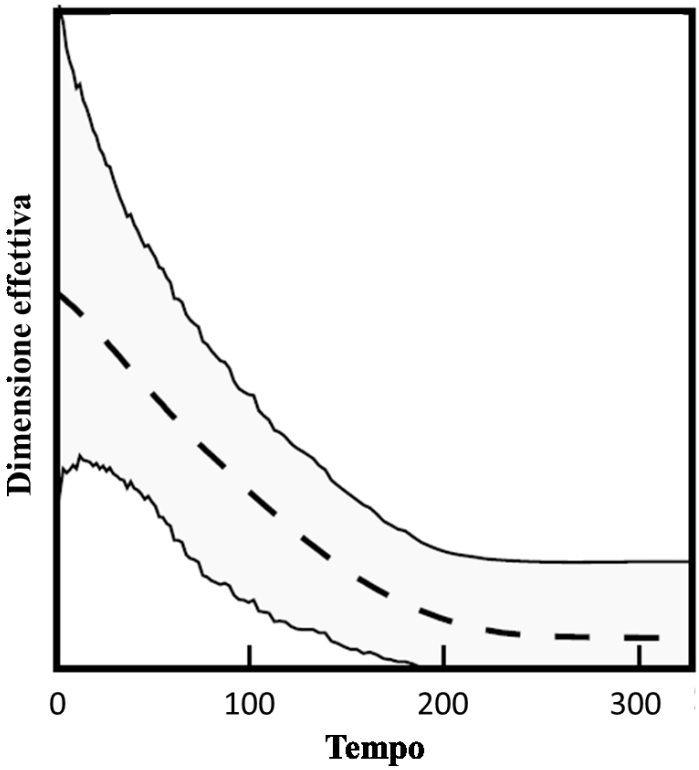

Figura1.3: Rappresentazione grafica di uno skyline plot. La dimensione effettiva (in ordinata) è visualizzata in funzione del tempo (in ascissa). Il tempo è misurato dal presente (0) fino al tempo dell'antenato comune più recente (TMRCA). La dimensione effettiva La linea tratteggiata rappresenta la mediana della ricostruzione mentre in grigio è evidenziato l'HPD.

In questo studio la dinamica della dimensione effettiva nel tempo è stata stimata con il

metodo “Extended Bayesian Skyline Plot” implementato nel software BEASTv1.6.1 (Drummond et al. 2002; Heled e Drummond 2008). Questo metodo, basato su un framework bayesiano accoppiato a Monte Carlo Markov Chain (MCMC), permette di stimare la genalogia degli individui simulati a partire dai dati di variabilità molecolare e ricostuire la funzione demografica nel tempo in un singolo passaggio. Inoltre, a differenza degli altri “Skyline Plot”, si possono analizzare contemporaneamente più loci indipendenti per stimare la storia demografica di una popolazione, riducendo in questo modo l’errore associato al Coalescente e, di conseguenza, l’errore nella stima. Ad ogni locus impiegato nell’analisi è possibile associare un fattore che tenga in considerazione la ploidia ed ereditabilità. In questo modo, ad esempio, è possibile tenere in considerazione che la dimensione effettiva di un locus autosomale trasmesso in maniera biparentale è quattro volte più grande di quella di un locus aploide mitocondriale. Inoltre, solo utilizzando l’ ”Extended Bayesian Skyline Plot” è possibile stimare il numero di cambiamenti demografici tramite “Bayesian Stochastic Variable Selection” (BSVS, Kuo e Mallick 1998). La presenza di molti intervalli di coalescenza corti può portare a un notevole aumento del “rumore di fondo” nella ricostruzione demografica perciò Strimmer e Pybus (2001) proposero di eliminare gli intervalli troppo corti raggruppandoli con i loro vicini in un numero *n* di gruppi specificato a priori. Con il BSVS è possibile selezionare l’*n* maggiormente supportato dai dati senza doverlo necessariamente specificare a priori. Se un solo gruppo viene selezionato, significa che i dati supportano uno scenario di popolazione costante nel tempo.

Framework di analisi

Ogni dataset simulato è stato convertito in un input file leggibile da BEAST attraverso il software BEAUTI disponibile nello stesso package. Lo stesso modello mutazionale utilizzato per la generazione dei dati simulati è stato impiegato nell’analisi (modello mutazionale HKY85, tasso di mutazione fissato a 1.25×10^{-7} mutazioni per sito per generazione, frequenza delle basi uguale e rapporto transizioni su trasversioni uguale a uno) in modo che i risultati non risentano dell’incertezza della stima dei parametri mutazionali. La distribuzione a priori del numero di cambiamenti demografici è stata definita come una distribuzione di Poisson con media ln(2) in modo da favorire nel 50% dei casi uno scenario di popolazione costante e nel restante 50% dei casi almeno un cambiamento demografico. La lunghezza della catena ha previsto 10 milioni di iterazioni con un campionamento dei parametri del modello ogni 1 000 iterazioni, ed il primo 10% della catena marcoviana totale è stato scartato (*burn-in*). I valori degli operatori che regolano il campionamento MCMC sono stati mantenuti invariati rispetto a quelli di default. Alla fine di ogni analisi è stato valutato se la convergenza è stata raggiunta, cioè se l’algoritmo ha campionato dalla

distribuzione a posteriori di ogni parametro. A questo scopo sono state calcolate due misure di convergenza per i parametri più importanti del modello: la likelihood, la distribuzione a priori globale (indica il campionamento complessivo da tutte le distribuzioni a priori di tutti i parametri), la distribuzione a posteriori complessiva (indica il campionamento complessivo da tutte le distribuzioni a posteriori di tutti i parametri) e numero di cambiamenti demografici. Come primo indice è stato calcolato il valore di "Effective Sample Size" (ESS) che rappresenta il numero di campionamenti indipendenti dalla distribuzione a posteriori stimata. L'algoritmo di campionamento di tipo MCMC, per sua natura, effettua dei campionamenti che sono correlati tra loro per cui il valore di ESS indica la qualità della stima della distribuzione a posteriori. Un ESS minore di 100 viene considerato in genere un valore basso e indicativo di problemi durante l'analisi. Come seconda indicazione di convergenza, è stato eseguito il test di Geweke. Questo test si basa sul principio che se la catena ha raggiunto la convergenza, la prima parte e l'ultima parte della catena avranno la stessa media e la loro differenza sarà distribuita in modo normale. Il comando geweke.diag disponibile nel package "CODA" per l'ambiente statistico R (R Development Core Team 2010) è stato utilizzato per condurre il test ed ottenere il p-value associato.

Per ognuna delle combinazioni demografiche studiate sono stati riassunti i dati di convergenza dei 100 dataset simulati calcolando per i quattro parametri considerati: la media di ESS, il numero di dataset che hanno un valore di ESS minore di 100 e la percentuale di dataset che hanno mostrato un p-value del Geweke test non significativo per un valore di α=0.05 (vedi Tabella1.9-11S per i risultati).

Le analisi dei dataset che hanno raggiunto la convergenza sono state utilizzate per ricostruire la dinamica demografica della popolazione nel tempo e rappresentarla in modo grafico per capire in maniera visiva gli scostamenti da uno scenario di popolazione costante.

Per ogni combinazione demografica è stata calcolata la probabilità (P_{ERR}) di favorire un modello demografico errato (cioè non una popolazione costante) verificando la percentuale di dataset (su 100 repliche) dove la distribuzione a posteriori del numero di cambiamenti demografici non conteneva lo 0 (un valore di 0 significa popolazione costante)(Heled e Drummond 2008). Inoltre è stata calcolata la mediana della distribuzione delle mode della distribuzione a posteriori del numero di cambiamenti demografici, in modo da riassumere il numero di cambiamenti supportato dai dati. Infine, per valutare l'intensità del cambiamento demografico è stato calcolato il rapporto tra la dimensione effettiva più antica (valore di Ne massimo nelle ultime 10 generazioni) e quella più recente (valore massimo tra la decima e la ventesima generazione). Valori superiori a uno indicano una riduzione demografica mentre valori inferiori a uno supportano uno scenario di espansione demografica avvenuta nel passato.

1.3 RISULTATI

Variabilità genetica presente negli scenari demografici studiati

I valori di variabilità genetica calcolati in 1 000 dataset simulati sono riportati in Tabella1.1S. La variabilità genetica intra popolazione si attesta su valori alti in ognuno degli scenari studiati (S medio circa 100, con i limiti dei percentili da 50 a 200) aumentando leggermente per livelli intensi di migrazione (m=10^{-2}, 10^{-1}) come atteso quando una popolazione riceve migranti. Il livello di divergenza tra le popolazioni moderne, è direttamente proporzionale al tempo della divergenza ancestrale e all'intensità della migrazione: i valori di Fst oscillano da una media di circa 0.5 quando la separazione è avvenuta 10 000 generazioni nel passato e la migrazione è bassa (m=10^{-5}), fino a valori prossimi allo 0 quando la divergenza è bassa (2 000 generazioni) e la migrazione è intensa (m=10^{-1},10^{-2}). Gli indici di neutralità (D di Tajima e Fs di Fu) non sembrano indicare cambiamenti demografici, producendo una frazione di dataset significativamente diversi da 0 estremamente bassa (% D max 0.10, % Fs max 0.17).

L'effetto della migrazione sulla ricostruzione demografica

Quando la migrazione tra le popolazioni è bassa (m=10^{-5}), le dinamiche ricostruite sono, come atteso, quelle tipiche di una popolazione di dimensioni costanti nel tempo (ad esempio Figura1.5, m=10^{-5}). Le curve ricostruite si sovrappongono completamente al controllo in tutte le combinazioni di parametri demografici e numero di loci (da Figura1.1S a Figura1.8S, m=10^{-5}). Inoltre, per questa intensità di migrazione, la P_{ERR} assume il valore 0 in tutti i casi analizzati (Figura1.3) e il rapporto *Na*/*Ne* rimane centrato su 1(Figura1.4, min 0.94 – max 1.03). Un aumento dell'intensità della migrazione influenza la ricostruzione demografica. Valori di m intermedi (10^{-4}, 10^{-3}) aumentano la probabilità di commettere un errore fino a valori di circa il 40% per valori di T_{MIG} maggiori di 150 generazioni e se vengono utilizzati più di un locus per l'analisi. Inoltre, se le popolazioni scambiano migranti da molto tempo (T_{MIG} > 150) e sono altamente differenziate (T_{MIG}=10 000), la curva ricostruita tende a piegarsi verso il basso per tempi prossimi al presente indicando una riduzione demografica in corso nella popolazione (vedi ad esempio Figura1.5 T_{MIG} = T_{DIV} = 10 000, Figura1.6-7-8S). Risultato confermato anche dal rapporto delle dimensioni effettive antiche e moderne che assume valori maggiori di uno, tipici di una riduzione demografica (Figura1.4, T_{DIV} = 10 000, T_{MIG} > 150). Inoltre, in queste condizioni di migrazione e tempi di divergenza/migrazione, più loci vengono utilizzati per l'analisi, più la distorsione nella

ricostruzione demografica risulta evidente (Figura1.3, Figura1.4, Figura1.5). Per alte intensità di migrazione (m=10^{-2}, 10^{-1}), si osservano due tipi di deformazioni della dinamica demografica ricostruita. Se la divergenza tra le popolazioni e l'inizio della migrazione sono entrambe antiche ($T_{MIG} = T_{DIV} = 10\ 000$), la dinamica ricostruita è quella di una popolazione di dimensioni costanti nel tempo ma doppie rispetto alle dimensioni reali (Figura1.5, Figura1.8S). Come atteso dall'osservazione delle dinamiche ricostruite, il valore di P_{ERR} è molto piccolo, variando dallo 0 al 10%, rispettivamente con 1 e 10 loci e il rapporto *Na*/*Ne* leggermente inferiore a 1 (Figura1.4). Negli altri casi invece, soprattutto per bassi livelli di divergenza (T_{DIV} = 2 000), la dinamica ricostruita è quella tipica di un'espansione demografica, dove l'inizio dell'espansione coincide con il reale tempo della divergenza tra le popolazioni (Figura1.5, Figura1.1S, Figura1.3S, Figura1.5S). Il rapporto *Na*/*Ne* tende a raggiungere valori di 0.5, indicando chiaramente un'espansione demografica che ha portato al raddoppiamento della dimensione antica (Figura1.4, T_{DIV} = 2 000, m=10^{-2} - 10^{-1}), e la P_{ERR} varia dal 20 al 90% quando si utilizzano 10 loci. Lo stesso andamento demografico si osserva in misura minore anche per popolazioni molto divergenti (T_{DIV} = 10 000) e tempi di migrazione recenti ($T_{MIG} \leq 5\ 000$). La media del rapporto tra *Na*/*Ne*, in questo caso, suggerisce un' espansione demografica, assumendo valori minori di uno (Figura1.4, $T_{MIG} \leq 5\ 000$, T_{DIV} = 10 000) ma l'ispezione grafica delle ricostruzioni demografiche indica che la distorsione demografica non è supportata dalla maggior parte delle ricostruzioni (Figura1.2S, Figura1.4S, Figura1.6S, Figura1.7S). Interessante notare che in generale, le analisi basate su un singolo locus raramente ottengono P_{ERR} maggiore di 0, cioè riescono rifiutare l'ipotesi di popolazione costante, a differenza delle analisi multilocus dove in alcuni casi lo scenario corretto viene rifiutato 90 volte su 100.

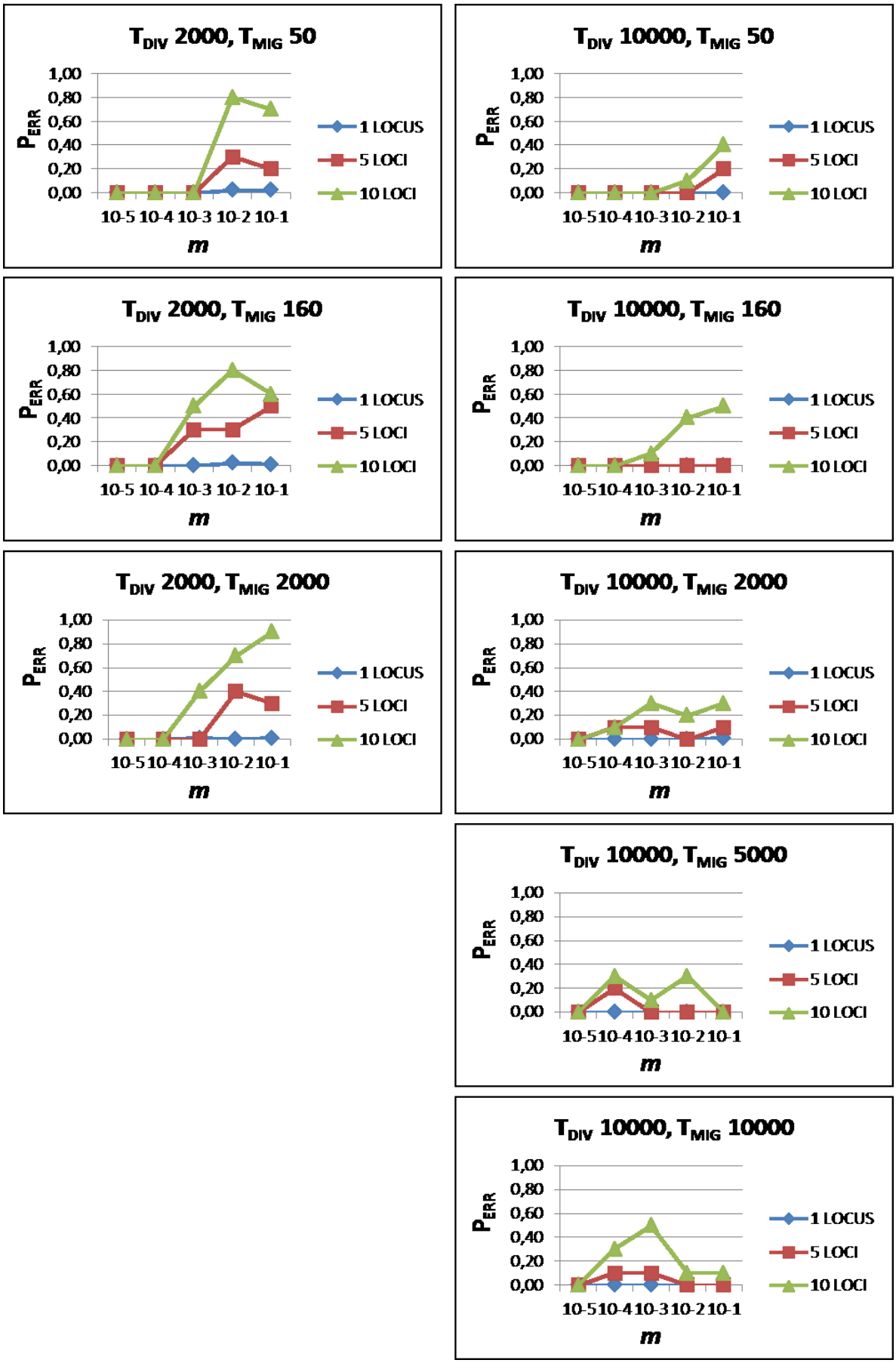

Figura 1.3: probabilità di identificare un cambiamento demografico (P_{ERR}) al variare dell'intensità della migrazione (*m*). Le linee di diversi colori corrispondono alla probabilità calcolata con 1, 5 o 10 loci.

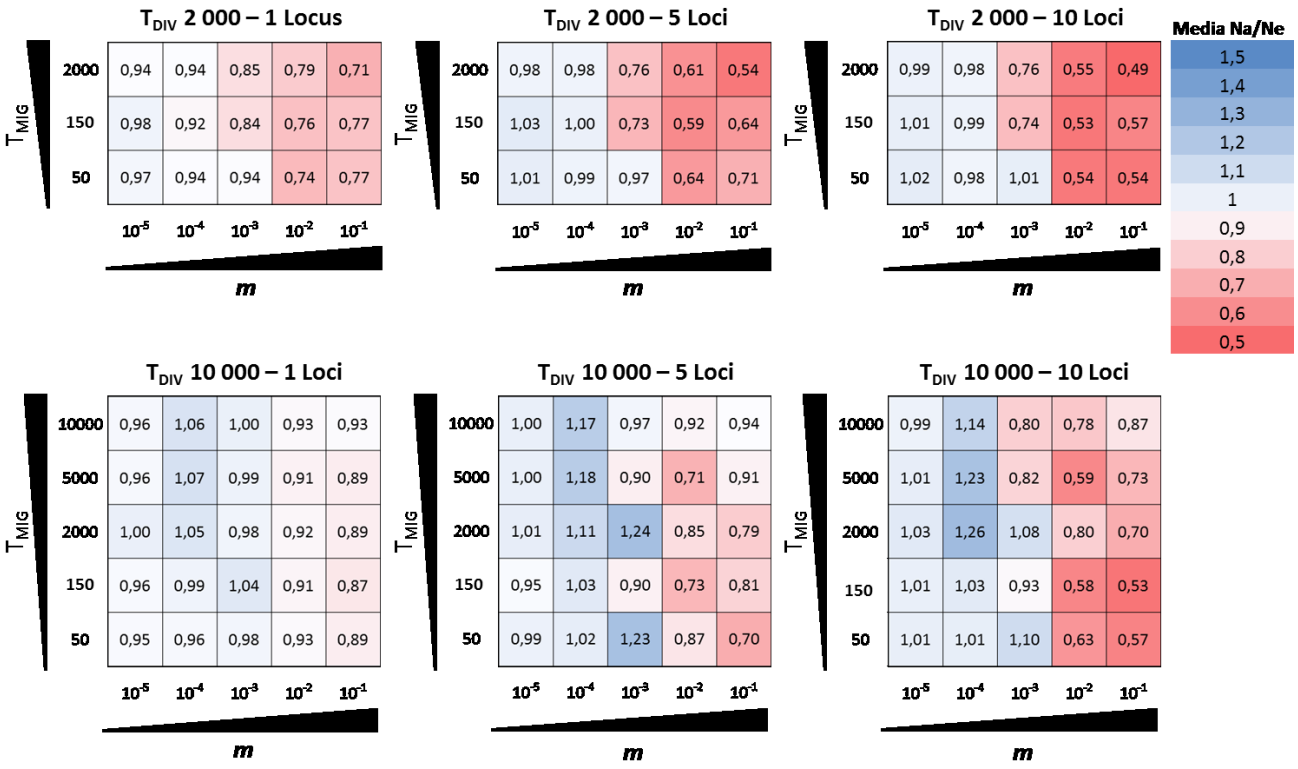

Figura 1.4: Media del rapporto tra la dimensione effettiva ancestrale (*Na*) e moderna (*Ne*). Valori superiori a uno (evidenziati in tonalità di blu) indicano una riduzione demografica, mentre valori inferiori a uno (evidenziati in tonalità di rosso) un'espansione demografica.

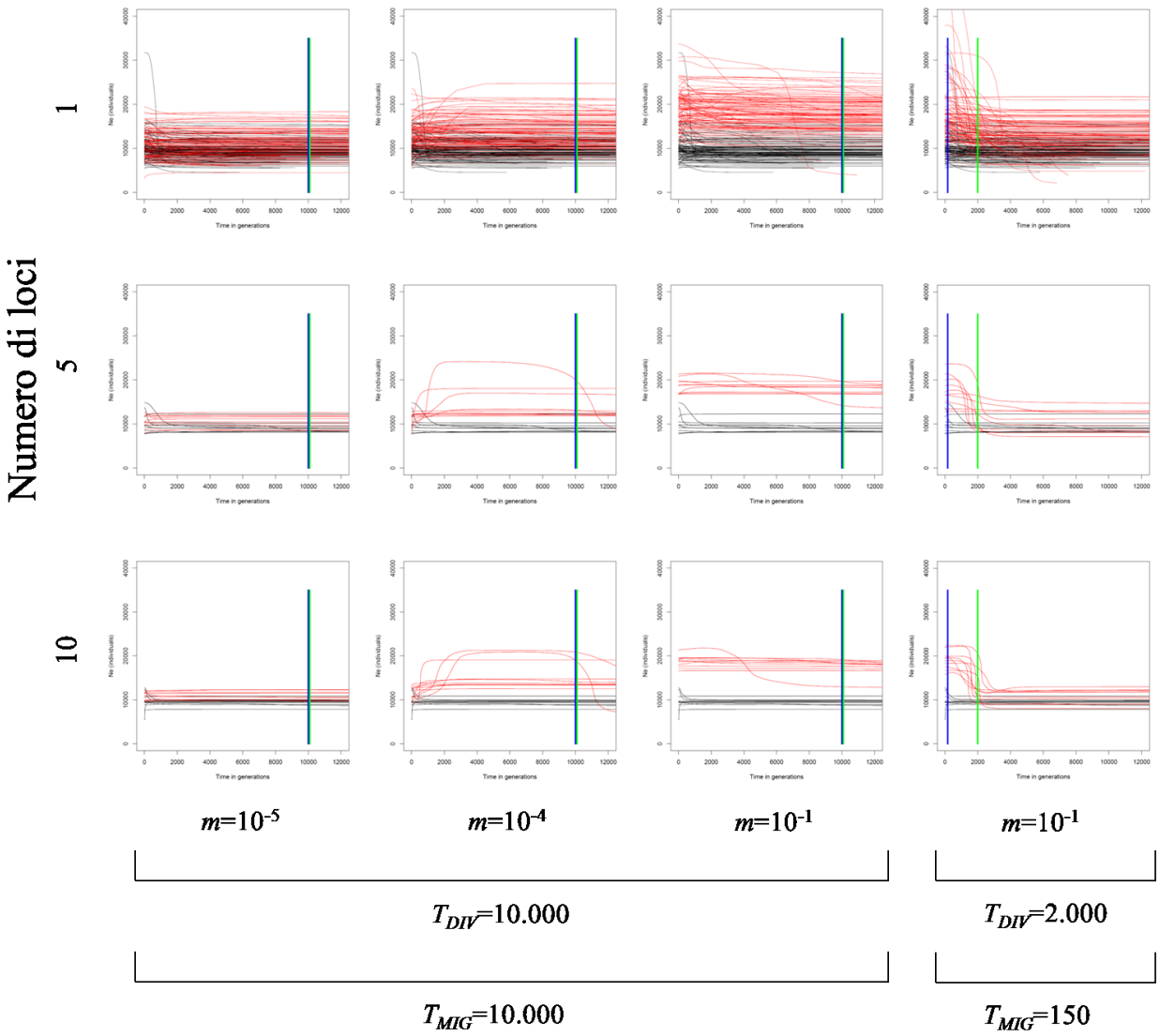

Figura 1.5: rappresentazione grafica riassuntiva dell'influenza della migrazione sulle ricostruzioni demografiche. In ogni grafico sono riportate le mediane degli skyline plot ricostruiti in assenza (linee in rosso in nero) e in presenza di migrazione. Il tempo di inizio della divergenza (T_{DIV}) e della migrazione (T_{MIG}) sono indicati graficamente rispettivamente dalla linea retta verde e blu.

1.4 DISCUSSIONE

L'accurata ricostruzione della storia demografica delle popolazioni è un passaggio chiave per rispondere a numerose domande della biologia evoluzionistica che vanno dalla gestione delle popolazioni a rischio di estinzione (Hansen et al. 2008; Valdiosera et al. 2008), all'inferenza sulle dinamiche epidemiologiche di virus e batteri (Rambaut et al. 2008), allo studio della divergenza tra popolazioni (Hey 2010). Alcuni metodi analitici, come lo Skyline Plot, permettono un'inferenza molto accurata della storia demografica delle popolazioni, però esiste la necessità di conoscere il loro potere e affidabilità quando si analizzano condizioni tipiche dei sistemi naturali reali. Gli studi di simulazione si sono rivelati uno strumento molto utile per raggiungere questo obbiettivo e per capire nel dettaglio il comportamento di un metodo quando alcune assunzioni vengono violate (Hey 2005; Strasburg e Rieseberg 2010). In questo studio è stato analizzato come la presenza di migrazione influisca sulla ricostruzione della dinamica di una popolazione di dimensione costante nel tempo quando il metodo assume di analizzare una popolazione isolata.

Il metodo si è dimostrato robusto a bassi livelli di migrazione (m=10^{-5}), ricostruendo in tutti i casi analizzati lo scenario corretto e non rifiutando mai lo scenario di popolazione costante. L'arrivo di pochi migranti non sembra influenzare il pattern di coalescenze attese nella popolazione da cui proviene il campione, non influendo perciò nella dinamica ricostruita (vedi Figura1.3, Figura1.4, Figura1.5 prima colonna). Livelli di migrazione intermedi (m=10^{-4}, 10^{-3}) possono avere diversi effetti sulla ricostruzione demografica. Se la popolazione scambia migranti per poche generazioni (T_{MIG}=50, 150), non si osservano effetti sulla dinamica ricostruita, anche se la popolazione non campionata è altamente divergente. Lo "Skyline Plot" perciò risulta essere robusto alla violazione di isolamento o per valori estremamente bassi di migrazione o quando questa è avvenuta in tempi molto recenti e per un periodo limitato nel tempo. Scenari con livelli di migrazione intermedi tra popolazioni altamente divergenti (T_{DIV}=10 000) e per lungo periodo (T_{MIG} >150 gen) sono caratterizzati da ricostruzioni demografiche con una dinamica tipica di una popolazione in declino (Figura1.5 colonna due, Figura 1.4). Come già osservato in altri studi (Hein et al. 2005; Chikhi et al. 2010), l'entrata nella popolazione di individui altamente divergenti può in alcuni casi riprodurre l'effetto di una popolazione strutturata al suo interno, introducendo lunghi rami interni nella genealogia ricostruita e pattern di coalescenze tipiche di una popolazione in declino. Popolazioni che scambiano centinaia di individui per generazione sono da considerarsi come un'unica grande popolazione panmittica. Come atteso, alti livelli di migrazione (m=10^{-2}, 10^{-3}), provocano due diversi tipi di distorsione della dinamica di popolazione ricostruita. Quando la divergenza tra le popolazioni è recente (T_{DIV}=2 000), procedendo indietro nel tempo la dinamica

ricostruita è formata da una prima parte con una dimensione effettiva di 20 000 individui, cioè doppia rispetto al vero e una seconda parte dopo il tempo di divergenza di dimensioni di 10 000 (Figura1.5 colonna tre). Questo andamento demografico visto nel normale scorrere del tempo è tipico di una espansione demografica che ha raddoppiato la dimensione effettiva circa 2 000 generazioni nel passato e compare anche per tempi di migrazione molto brevi (ad esempio T_{MIG}=50). Più loci vengono utilizzati per l'analisi più questo fenomeno sembra essere chiaro, mentre utilizzando un singolo locus si hanno molti casi in cui l'andamento ricostruito è tipico di una popolazione costante. Anche quando la divergenza tra le popolazioni è più antica (T_{DIV}=10 000), si osserva lo stesso fenomeno (dove il tempo della divergenza coincide con il tempo di una finta espansione demografica), ma solo per tempi di migrazione brevi (T_{MIG} < 2 000) (Figura1.2S, Figura1.4S). Questo fenomeno è da tenere in considerazione quando si analizzano popolazioni reali che, ad esempio, abbiano recentemente colonizzato un nuovo habitat ma continuino a scambiare migranti con la popolazione di origine. In questo caso, l'iniziale evento di divergenza potrebbe essere confuso con un segnale di espansione demografica non realmente avvenuto. Infine, quando la migrazione avviene ad alti livelli per molte generazioni, il segnale di divergenza tra le popolazioni viene perso e la dinamica ricostruita è quella di un'unica popolazione costante nel tempo ma di dimensioni doppie rispetto al vero. Nei casi in cui sia possibile, sembra dunque essere necessario valutare il livello migrazione da e verso la popolazione oggetto di studio, prima di procedere alla ricostruzione demografica. Nel caso in cui ci siano evidenti segnali di flusso genico in corso, la ricostruzione demografica deve essere eseguita sapendo che potrebbe essere altamente influenzata, o addirittura evitata. Ad esempio nel lavoro pubblicato da Liao et al. (2010), gli autori hanno studiato come l'ultimo periodo glaciale abbia influenzato la storia demografica di quattro popolazioni di *Cinnamomum kanehirae* nell'isola di Taiwan. La dinamica demografica è stata ricostruita per ognuna di esse, sebbene siano presenti evidenti segnali di flusso genico antico e recente tra le popolazioni, e di conseguenza le assunzioni del modello alla base del metodo statistico siano palesemente violate. In questo caso, in accordo con i risultati ottenuti, la presunta espansione demografica, coincidente al ritiro dei ghiacci successivo al Last Glacial Maximum, potrebbe essere spiegata da uno scambio di migranti tra le popolazioni piuttosto che da un aumento della dimensione effettiva delle popolazioni. In conclusione, questo studio, sebbene limitato a specifiche condizioni di studio, suggerisce che la migrazione può influenzare profondamente le ricostruzioni demografiche prodotte con l'EBSP, e che l'interpretazione dei grafici ottenuti da dati genetici reali dovrebbe tenere in considerazione l'effetto i livelli di migrazione possono produrre per evitare conclusioni non corrette.

1.5 MATERIALI SUPPLEMENTARI

Tabella 1.1S: Indici di variabilità genetica intra e tra popolazioni negli scenari demografici simulati. I valori riferiti ai siti polimorfici e Fst si riferiscono alla media, 2,5% e 97,5% della distribuzione delle statistiche calcolate in 1.000 simulazioni coalescenti. Per la D di Tajima (D) e Fs di Fu (Fs) è stata riportatata la percentuale di volte in cui la statistica è significativamente maggiore (%pos) o minore (%neg) di 0.

				Siti polimorfici			Fst			D		Fs	
Modello	**T_{MIG}**	**T_{DIV}**	**m**	**Media**	**0.025**	**0.975**	**Media**	**0.025**	**0.975**	**%pos**	**%neg**	**%pos**	**%neg**
1	-	-	-	94,3	52,0	156,1	-	-	-	0,04	0,04	0,06	0,03
25	50	2000	10^{-5}	104,2	56,0	187,2	0,160	0,014	0,463	0,04	0,03	0,03	0,04
22	50	2000	10^{-4}	97,5	52,0	175,3	0,153	0,021	0,399	0,02	0,03	0,04	0,02
23	50	2000	10^{-3}	101,2	53,0	162,2	0,131	0,015	0,356	0,01	0,04	0,01	0,08
34	50	2000	10^{-2}	124,9	73,0	204,0	0,024	-0,023	0,143	0,03	0,04	0,01	0,14
24	50	2000	10^{-1}	127,5	73,0	223,0	0,001	-0,027	0,060	0,06	0,02	0,06	0,03
9	150	2000	10^{-5}	99,1	47,0	221,2	0,155	0,008	0,387	0,02	0,03	0,04	0,02
2	150	2000	10^{-4}	98,6	53,0	169,0	0,137	0,025	0,356	0,02	0,06	0,01	0,05
3	150	2000	10^{-3}	107,0	61,0	184,4	0,107	0,002	0,318	0,01	0,08	0,01	0,13
35	150	2000	10^{-2}	118,3	75,0	176,0	0,000	-0,027	0,065	0,02	0,04	0,01	0,10
7	150	2000	10^{-1}	127,5	72,9	235,6	0,004	-0,026	0,076	0,03	0,05	0,04	0,05
8	2000	2000	10^{-5}	99,6	49,0	165,2	0,186	0,018	0,565	0,03	0,05	0,04	0,05
4	2000	2000	10^{-4}	106,6	55,0	202,1	0,118	0,003	0,326	0,01	0,05	0,02	0,08
5	2000	2000	10^{-3}	114,0	67,9	202,2	0,023	-0,021	0,140	0,01	0,04	0,01	0,12
36	2000	2000	10^{-2}	119,8	76,9	189,2	-0,001	-0,025	0,065	0,01	0,06	0,01	0,17
6	2000	2000	10^{-1}	117,4	74,9	189,3	0,001	-0,025	0,066	0,03	0,05	0,03	0,04
29	50	10000	10^{-5}	92,8	45,9	167,2	0,504	0,266	0,801	0,05	0,04	0,05	0,04
26	50	10000	10^{-4}	100,2	53,0	167,1	0,458	0,224	0,812	0,01	0,07	0,02	0,02
27	50	10000	10^{-3}	124,4	64,0	221,2	0,401	0,157	0,726	0,01	0,02	0,01	0,04
37	50	10000	10^{-2}	152,3	96,9	210,2	0,070	-0,017	0,202	0,01	0,02	0,02	0,03
28	50	10000	10^{-1}	159,2	97,0	249,1	0,000	-0,026	0,043	0,02	0,04	0,05	0,03
20	150	10000	10^{-5}	101,2	53,0	169,2	0,468	0,191	0,727	0,03	0,07	0,05	0,03
11	150	10000	10^{-4}	111,0	50,0	183,2	0,420	0,161	0,761	0,01	0,06	0,02	0,01
14	150	10000	10^{-3}	142,4	92,0	252,1	0,295	0,103	0,551	0,01	0,05	0,01	0,03
38	150	10000	10^{-2}	161,8	109,0	227,1	0,000	-0,026	0,040	0,02	0,02	0,01	0,03
17	150	10000	10^{-1}	157,1	106,0	227,1	-0,001	-0,027	0,054	0,01	0,07	0,03	0,03
19	2000	10000	10^{-5}	100,5	53,0	183,1	0,446	0,231	0,767	0,03	0,05	0,05	0,01
10	2000	10000	10^{-4}	130,9	74,0	206,4	0,274	0,063	0,614	0,01	0,03	0,03	0,02
13	2000	10000	10^{-3}	146,3	90,0	229,2	0,033	-0,015	0,124	0,01	0,02	0,01	0,02
39	2000	10000	10^{-2}	157,9	102,9	224,2	0,002	-0,024	0,053	0,01	0,02	0,01	0,05
16	2000	10000	10^{-1}	153,3	90,9	249,1	0,001	-0,026	0,061	0,05	0,03	0,07	0,04
33	5000	10000	10^{-5}	105,2	46,0	204,1	0,417	0,150	0,701	0,02	0,04	0,05	0,01
30	5000	10000	10^{-4}	131,9	78,0	221,1	0,195	0,032	0,463	0,01	0,01	0,03	0,02
31	5000	10000	10^{-3}	147,1	99,0	210,1	0,018	-0,018	0,111	0,01	0,02	0,01	0,06
40	5000	10000	10^{-2}	152,9	95,0	240,6	0,008	-0,026	0,067	0,01	0,02	0,01	0,04
32	5000	10000	10^{-1}	152,9	91,0	268,1	0,002	-0,025	0,060	0,03	0,05	0,02	0,03
21	10000	10000	10^{-5}	103,8	50,0	171,1	0,419	0,161	0,747	0,01	0,02	0,07	0,02
12	10000	10000	10^{-4}	130,9	73,1	225,0	0,184	0,025	0,494	0,01	0,03	0,04	0,01
15	10000	10000	10^{-3}	151,5	96,0	231,3	0,022	-0,019	0,113	0,01	0,04	0,01	0,04
41	10000	10000	10^{-2}	151,9	100,0	218,4	0,010	-0,024	0,111	0,01	0,03	0,03	0,06
18	10000	10000	10^{-1}	158,4	100,0	229,1	0,002	-0,026	0,091	0,03	0,04	0,04	0,02

Figura 1.1S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 50 (barra verticale blu), T_{DIV} 2 000 (barra verticale verde).

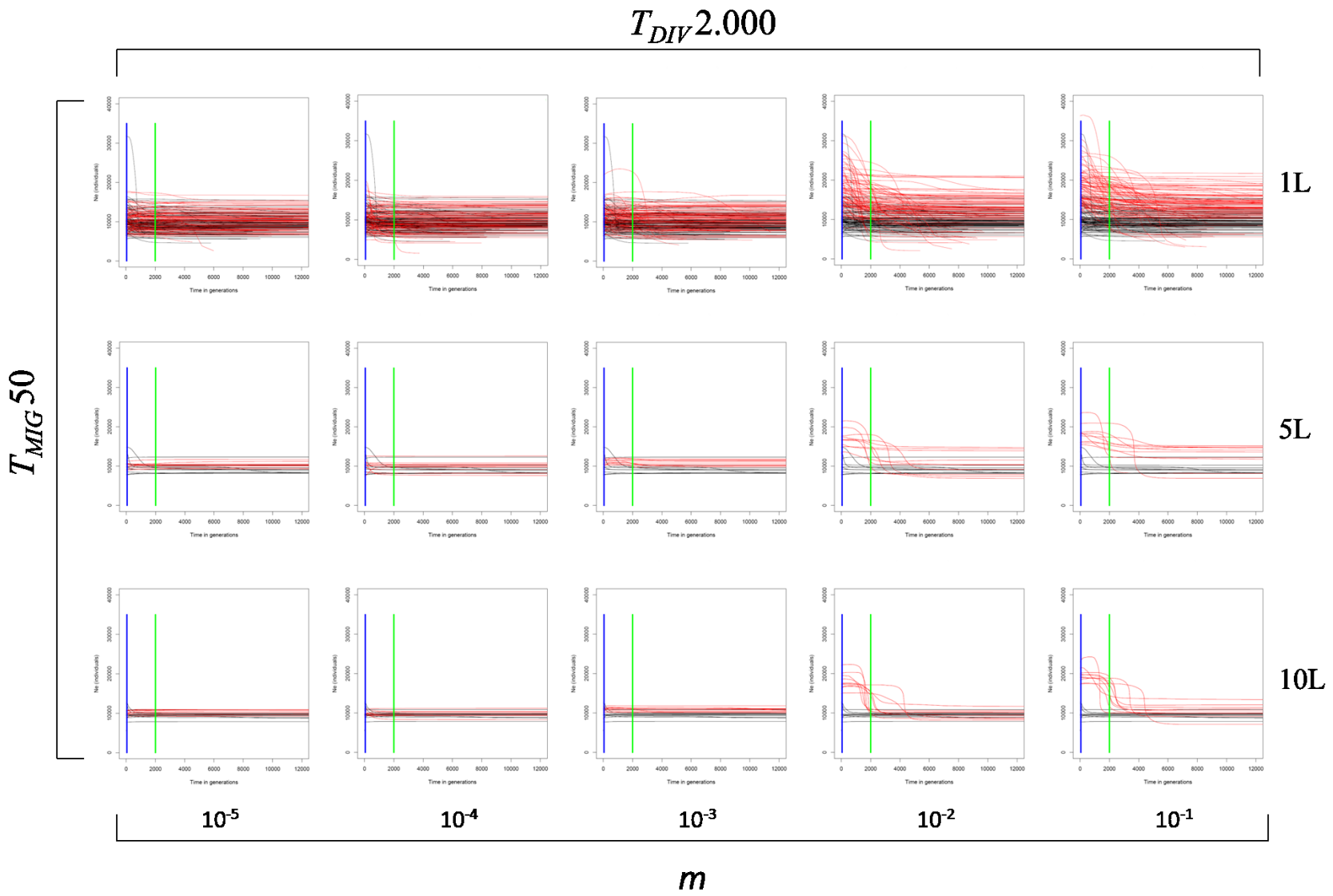

Figura 1.2S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 50 (barra verticale blu), T_{DIV} 10 000 (barra verticale verde).

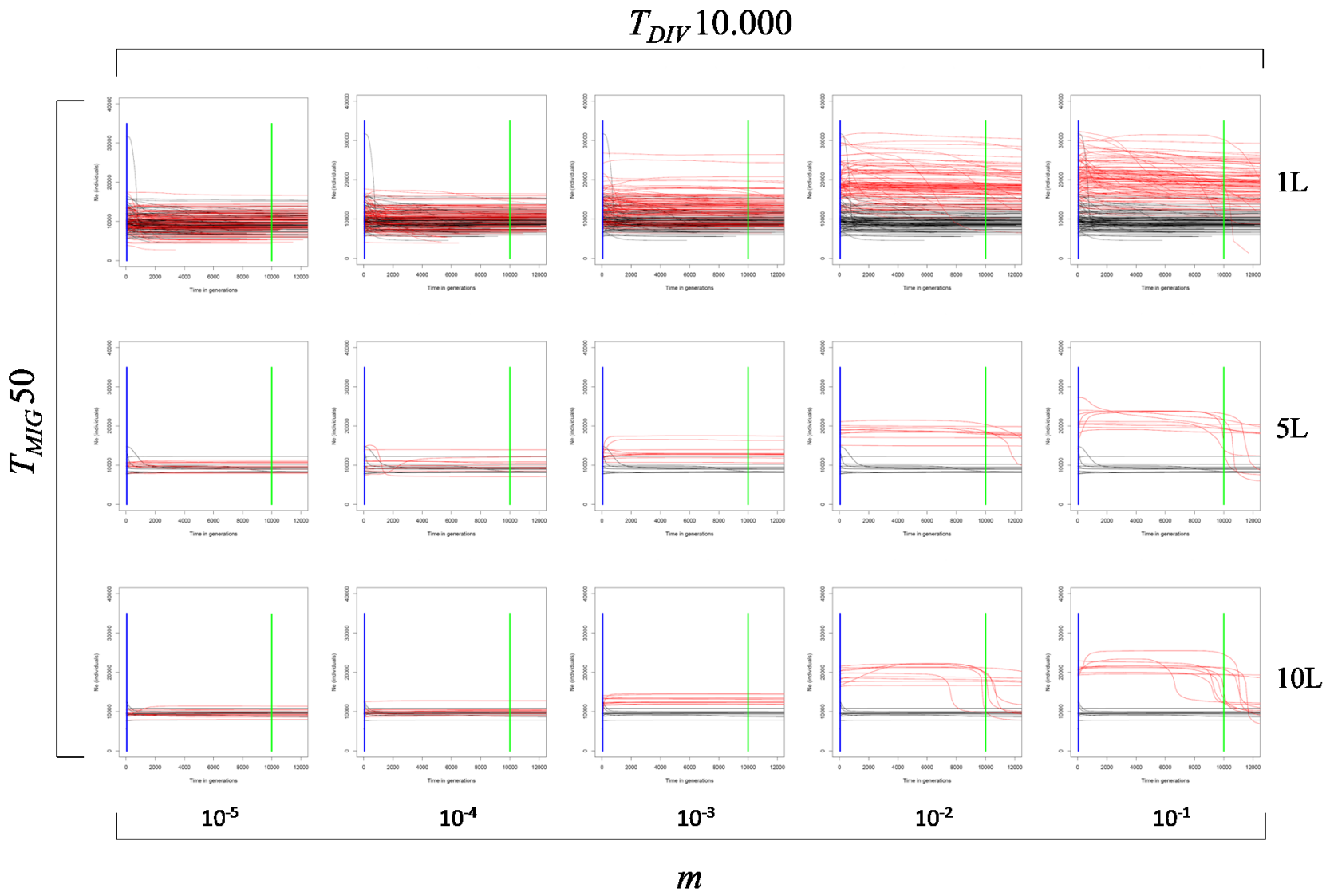

Figura 1.3S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 150 (barra verticale blu), T_{DIV} 2 000 (barra verticale verde).

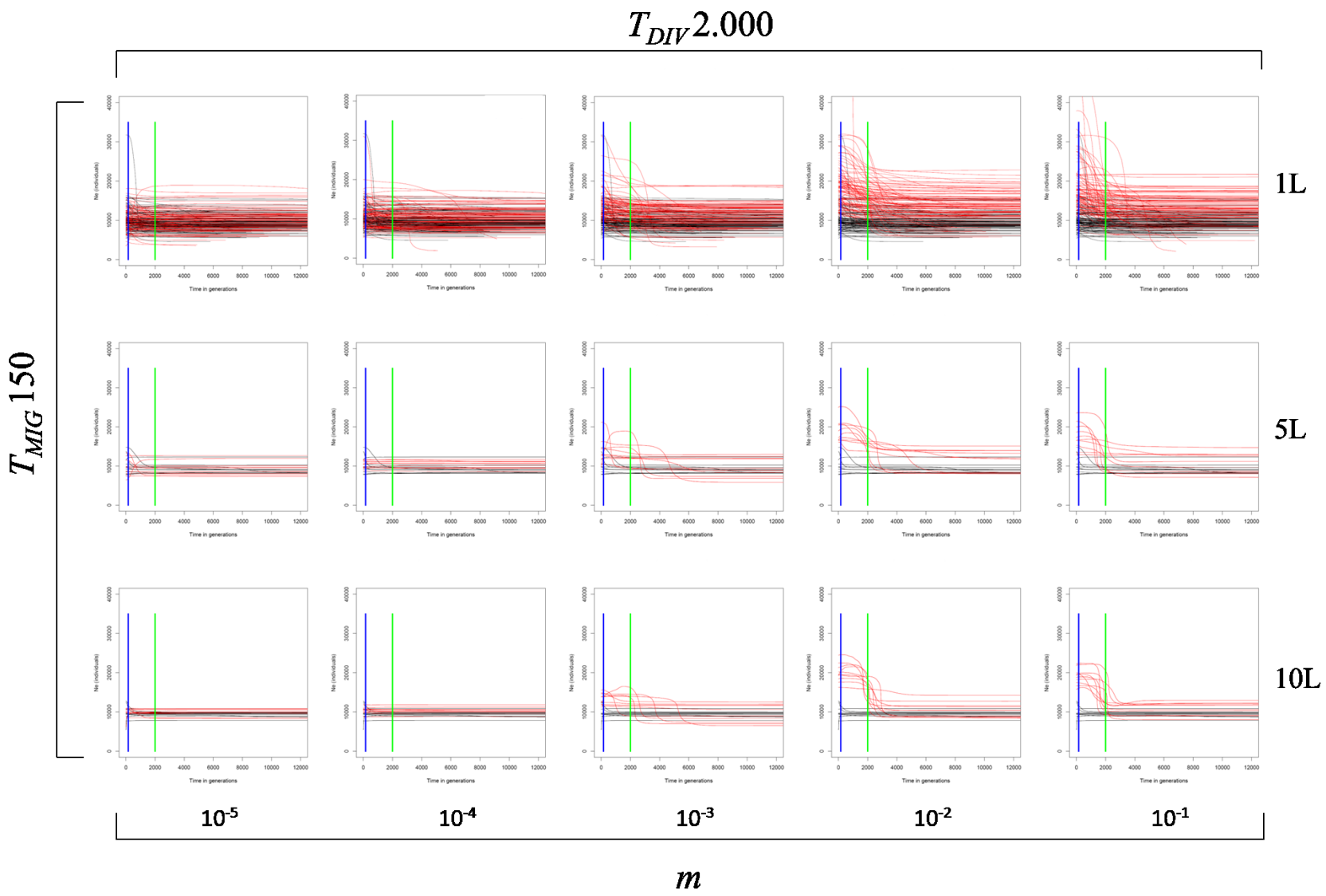

Figura 1.4S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 150 (barra verticale blu), T_{DIV} 10 000 (barra verticale verde).

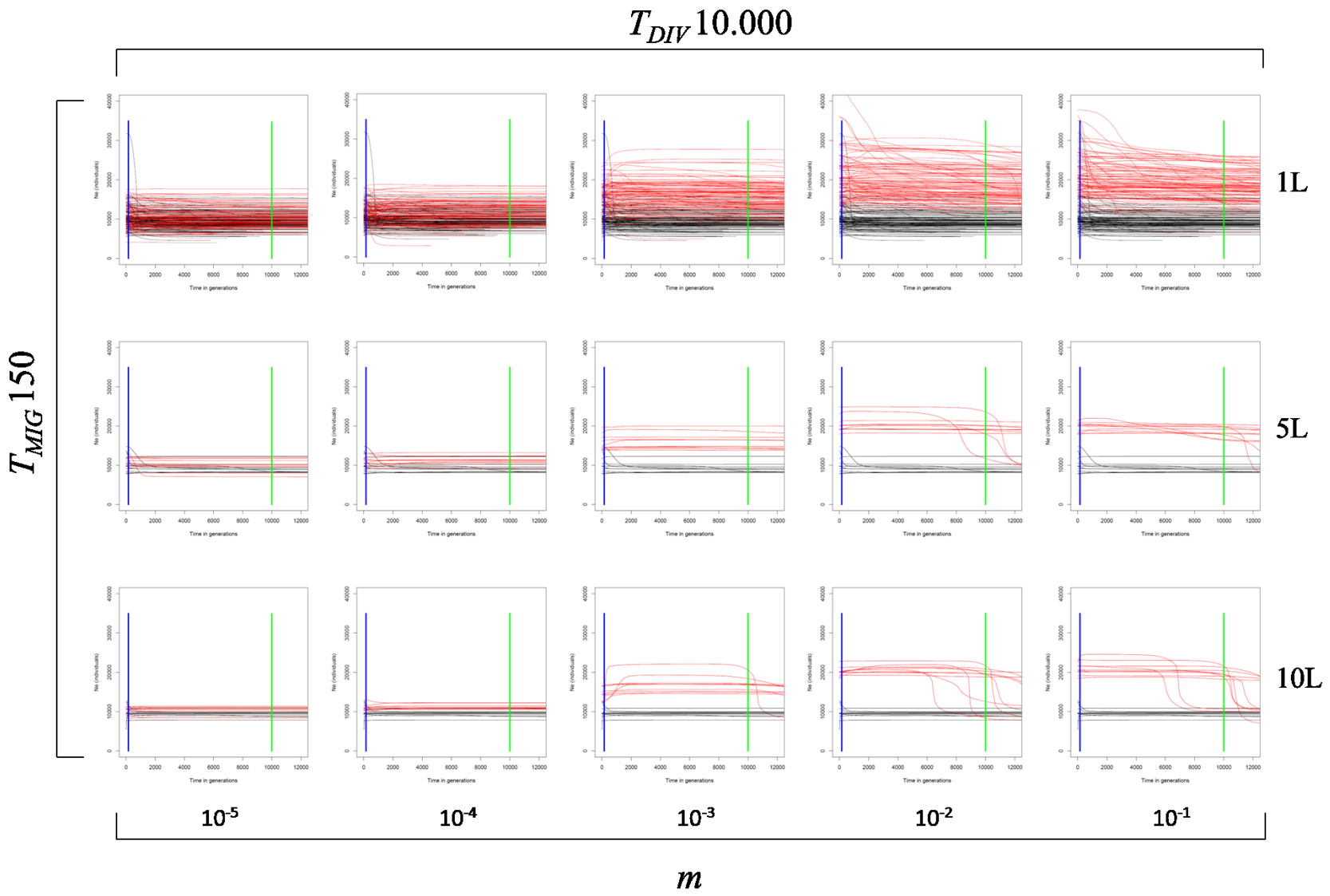

Figura 1.5S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 2 000 (barra verticale blu), T_{DIV} 2 000 (barra verticale verde).

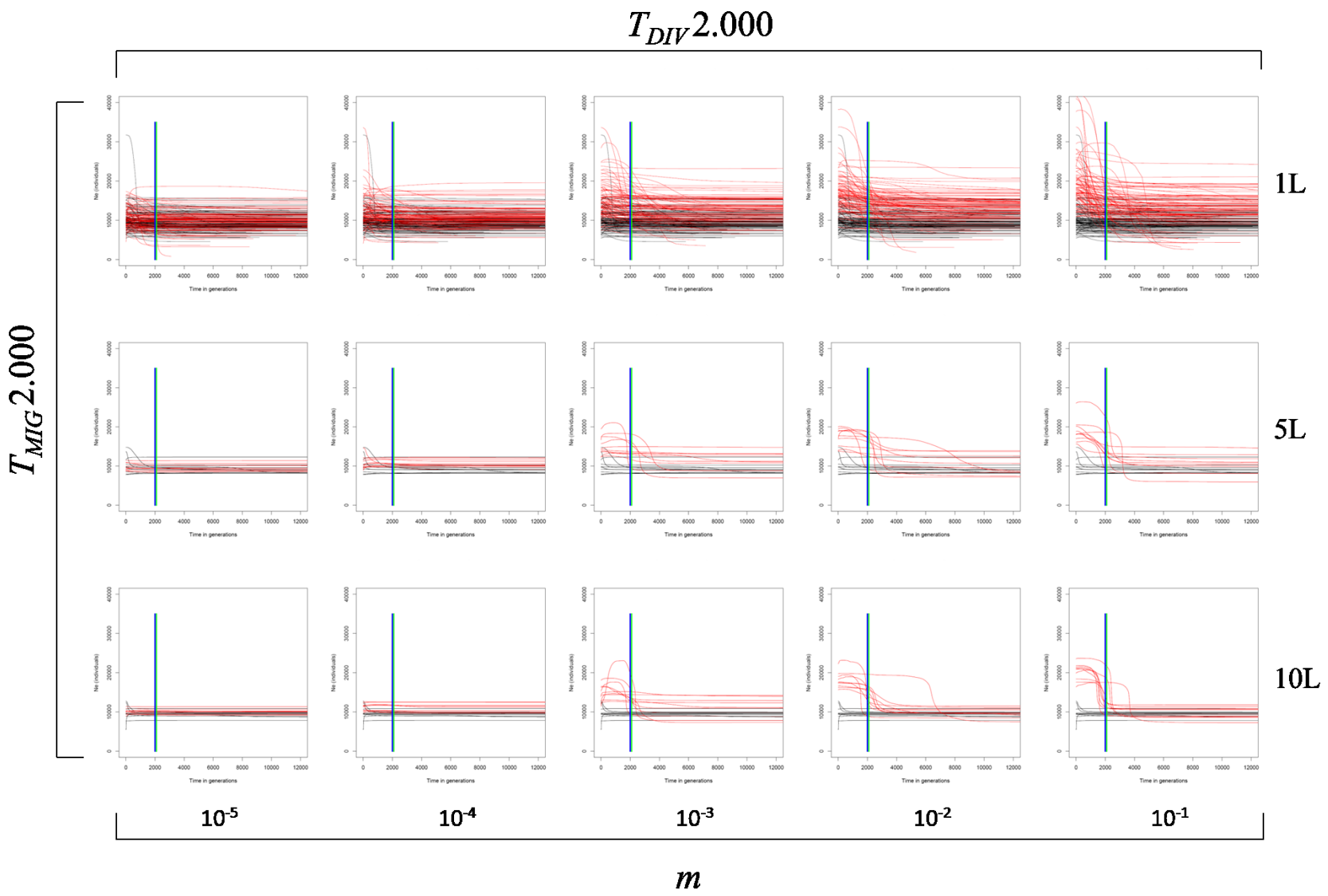

Figura 1.6S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 2 000 (barra verticale blu), T_{DIV} 10 000 (barra verticale verde).

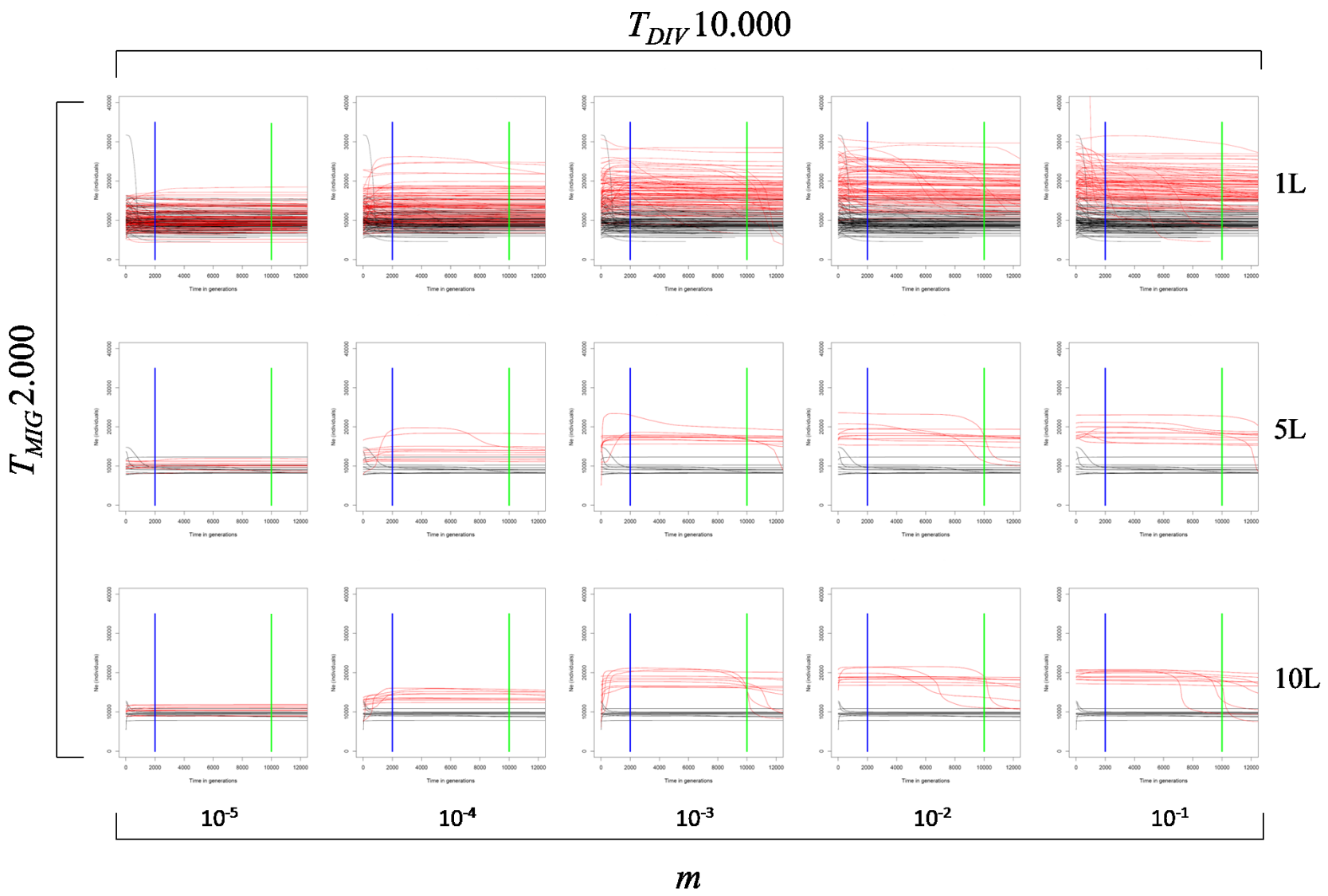

Figura 1.7S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 5 000 (barra verticale blu), T_{DIV} 10 000 (barra verticale verde).

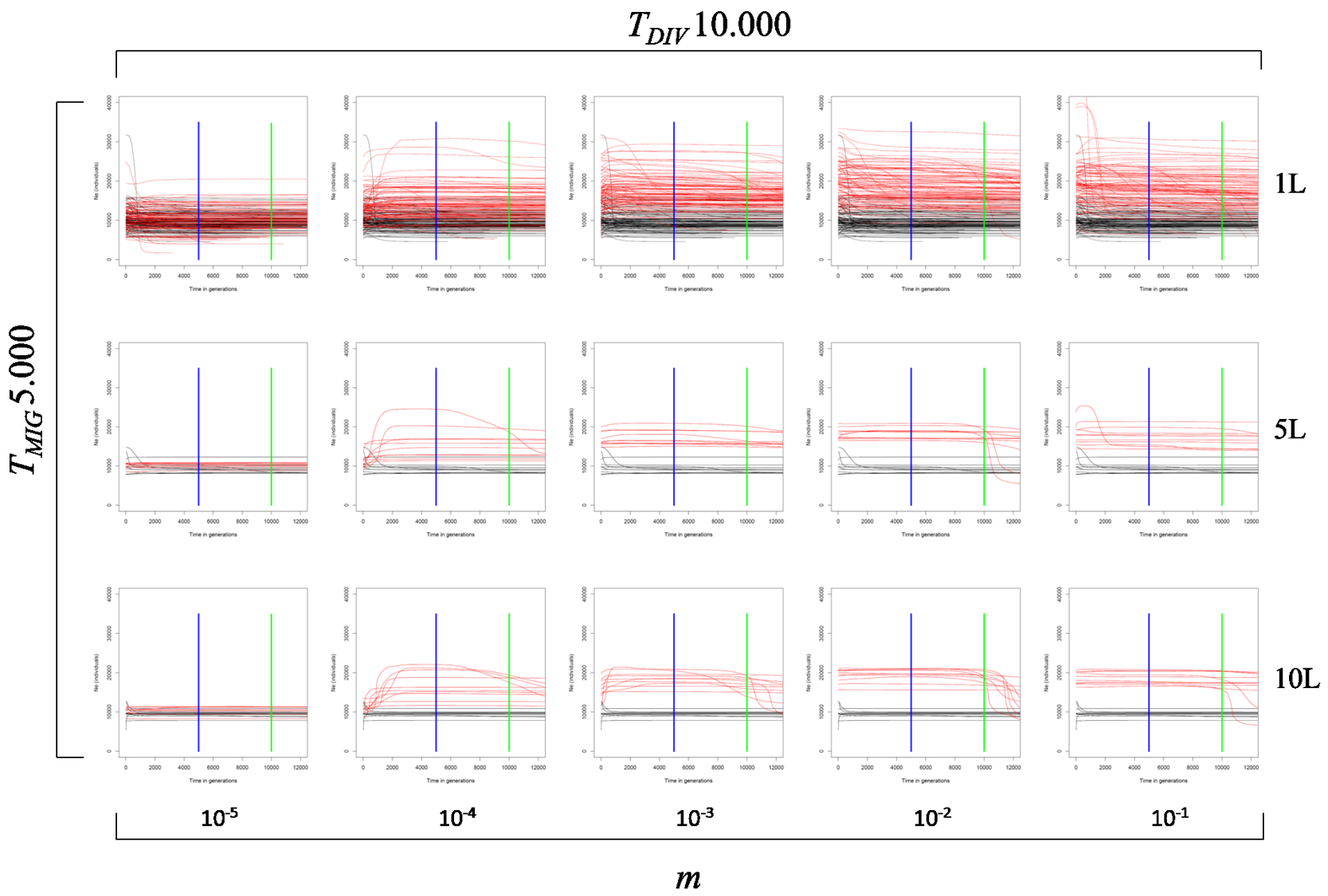

Figura 1.8S: Skyline plot ricostruiti in assenza (linee in nero) o in presenza (linee in rosso) di migrazione a differenti intensità, utilizzando 1, 5 e 10 loci. T_{MIG} 10 000 (barra verticale blu), T_{DIV} 10 000 (barra verticale verde).

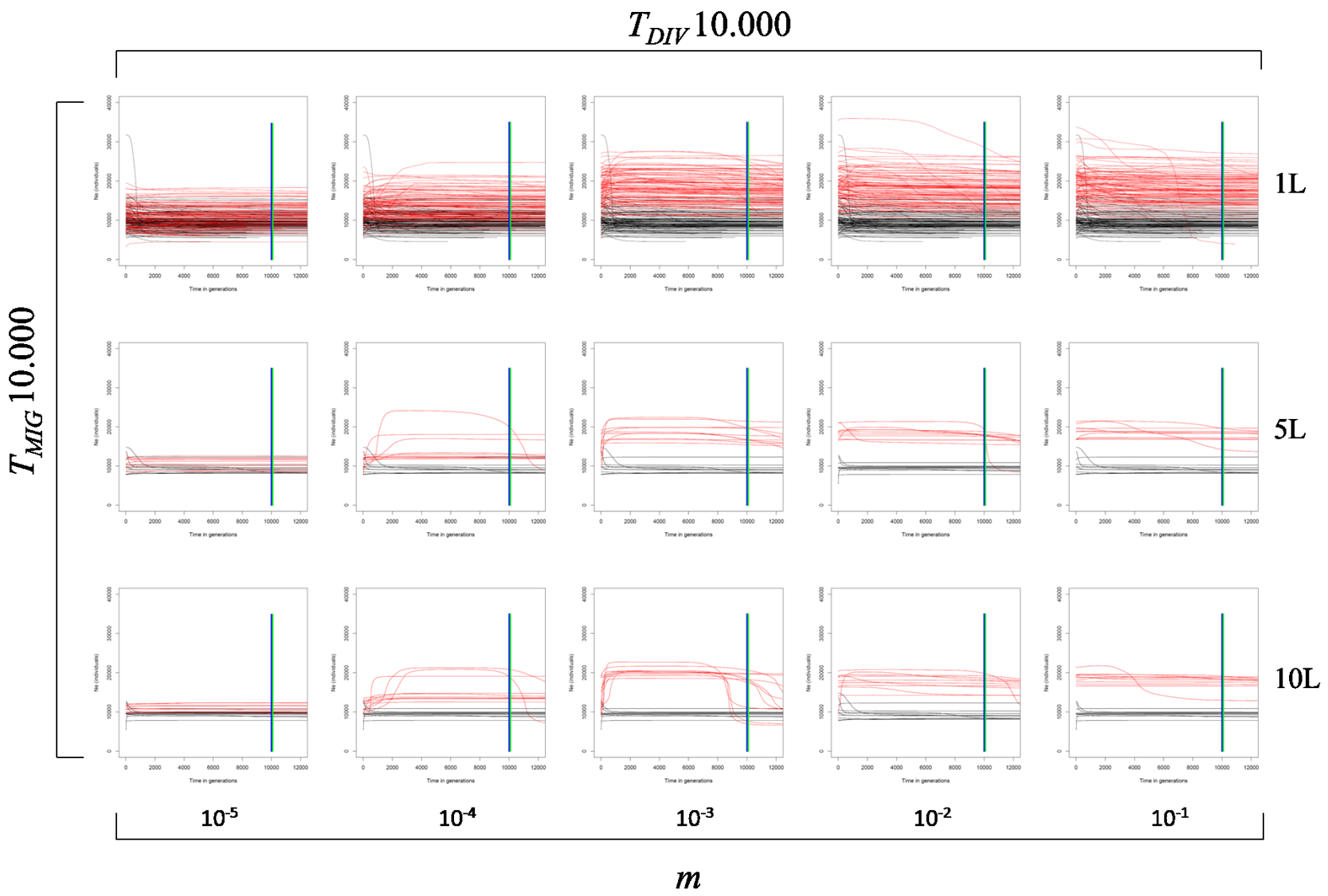

Tabella1.9S: indici di convergenza delle analisi effettuate usando un locus. Gli indici di convergenza sono stati calcolati sulla distribuzione a posteriori totale (Post), sulla distribuzione a priori totale (Pri), sulla likelihood (Lik) e sulla distribuzione a posteriori del numero di cambiamenti demografici (N° C).

	Media ESS				N° ESS < 100				% Geweke non significativi			
Modello	Post	Pri	Lik	N° C	Post	Pri	Lik	N° C	Post	Pri	Lik	N° C
1	183	182	2584	1821	64	67	0	0	0.85	0.86	0.95	0.92
2	371	367	5579	3912	4	4	0	0	0.88	0.88	0.94	0.95
3	376	371	5736	3957	4	4	0	0	0.87	0.87	0.91	0.89
4	369	366	5632	3895	3	2	0	0	0.91	0.91	0.95	0.93
5	374	371	5635	3989	4	3	0	0	0.83	0.82	0.95	0.89
6	416	414	5714	4015	3	2	0	0	0.90	0.90	0.97	0.90
7	378	375	5826	4050	6	7	0	0	0.87	0.88	0.93	0.92
8	379	372	5303	3850	0	0	0	0	0.86	0.85	0.98	0.95
9	380	375	5502	4096	1	2	0	0	0.91	0.91	0.93	0.93
10	352	349	5369	3898	0	1	0	0	0.85	0.85	0.95	0.95
11	362	357	5498	3804	3	3	0	0	0.87	0.88	0.89	0.90
12	341	337	5192	3762	4	4	0	0	0.86	0.87	0.95	0.92
13	346	342	5377	3668	4	4	0	0	0.86	0.86	0.96	0.95
14	338	334	5319	3899	4	4	0	0	0.88	0.88	0.93	0.91
15	334	329	5289	3802	4	4	0	0	0.88	0.88	0.94	0.93
16	335	333	5707	4045	4	4	0	0	0.91	0.91	0.97	0.90
17	336	332	5796	4120	6	5	0	0	0.88	0.88	0.92	0.95
18	335	333	5710	4102	3	2	0	0	0.89	0.89	0.98	0.89
19	373	369	5426	3709	2	2	0	0	0.88	0.89	0.93	0.94
20	355	353	5604	4072	2	2	0	0	0.90	0.89	0.93	0.94
21	359	356	5664	4116	2	2	0	0	0.86	0.87	0.92	0.94
22	378	378	5516	3915	3	2	0	0	0.90	0.90	0.94	0.91
23	379	375	5603	3938	3	4	0	0	0.92	0.92	0.97	0.91
24	376	373	5860	3949	1	2	0	0	0.87	0.87	0.95	0.91
25	362	359	5553	3982	3	2	0	0	0.88	0.88	0.94	0.93
26	375	370	5406	3875	2	3	0	0	0.88	0.88	1.00	0.93
27	337	336	5654	4128	3	3	0	0	0.87	0.87	0.97	0.96
28	343	339	5762	4175	4	3	0	0	0.91	0.91	0.98	0.92
29	370	366	5564	3927	4	4	0	0	0.86	0.87	0.96	0.94
30	347	342	5311	3650	2	3	0	0	0.89	0.90	0.94	0.91
31	315	311	5309	3669	8	11	0	0	0.88	0.88	0.96	0.90
32	350	346	5731	4106	5	5	0	0	0.91	0.91	0.92	0.94
33	369	366	5333	3751	3	3	0	0	0.84	0.84	0.92	0.93
34	401	398	5617	3944	2	2	0	0	0.88	0.87	0.94	0.94
35	394	391	5794	4155	3	3	0	0	0.90	0.90	0.97	0.94
36	383	381	5860	4049	4	4	0	0	0.83	0.83	0.92	0.94
37	323	318	5592	4008	7	7	0	0	0.89	0.88	0.93	0.91
38	333	330	5738	4057	2	2	0	0	0.91	0.91	0.96	0.95
39	321	317	5666	4065	8	9	0	0	0.87	0.89	0.96	0.97
40	327	324	5758	3925	7	7	0	0	0.91	0.92	0.98	0.95
41	329	328	5769	4252	7	8	0	0	0.90	0.90	0.94	0.93

Tabella1.10S: indici di convergenza delle analisi effettuate usando 5 loci. Gli indici di convergenza sono stati calcolati sulla distribuzione a posteriori totale (Post), sulla distribuzione a priori totale (Pri), sulla likelihood (Lik) e sulla distribuzione a posteriori del numero di cambiamenti demografici (N° C).

	Media ESS				N° ESS < 100				% Geweke non significativi			
Modello	Post	Pri	Lik	N° C	Post	Pri	Lik	N° C	Post	Pri	Lik	N° C
1	2107	2873	1717	1289	0	0	0	0	0.90	0.90	0.90	0.90
2	2547	3550	1881	1784	0	0	0	0	0.80	0.70	0.90	0.90
3	2533	3534	1643	1510	1	1	0	0	1.00	0.90	1.00	0.90
4	1796	2370	1487	1234	1	1	0	0	0.90	0.90	0.90	0.80
5	1610	2172	1585	1092	1	0	0	0	0.80	0.90	0.80	0.60
6	2937	4098	1882	1778	0	0	0	0	1.00	1.00	1.00	0.90
7	3182	4312	1830	2026	0	0	0	0	1.00	1.00	0.90	1.00
8	2434	3387	1746	1467	0	0	0	0	0.90	0.90	1.00	0.90
9	2431	3244	1768	1407	0	0	0	0	1.00	0.90	1.00	1.00
10	1412	1798	1503	935	1	2	0	1	0.90	0.90	0.90	0.90
11	1584	2186	1680	1151	0	0	0	0	0.80	0.80	1.00	0.70
12	1087	1729	1253	1111	2	2	0	0	0.90	0.90	1.00	1.00
13	923	1432	1113	846	2	3	0	0	0.80	0.80	0.80	0.70
14	1989	2794	1678	1485	0	0	0	0	1.00	1.00	1.00	0.90
15	1455	1897	1352	1113	1	1	0	0	0.80	0.70	1.00	0.90
16	2402	3788	1772	1540	0	0	0	0	0.80	0.80	1.00	0.80
17	2696	3919	1898	2232	0	0	0	0	1.00	1.00	1.00	0.90
18	2366	3660	1797	1657	0	0	0	0	1.00	0.90	1.00	0.90
19	1707	2146	1705	1192	0	0	0	0	0.90	0.90	1.00	0.90
20	2177	3184	1611	1388	0	0	0	0	0.90	0.90	0.90	0.80
21	1964	2679	1676	1554	0	0	0	0	0.70	0.80	0.90	0.70
22	2309	3374	1615	1428	0	0	0	0	1.00	1.00	1.00	0.90
23	2247	3199	1679	1445	1	1	0	0	0.80	0.80	0.90	0.90
24	3030	4413	1906	1476	0	0	0	0	0.80	0.90	0.90	1.00
25	2105	2783	1892	1459	0	0	0	0	0.90	0.90	0.90	0.90
26	2047	2838	1726	1314	0	0	0	0	0.80	0.80	0.90	0.90
27	1550	2209	1500	1402	1	1	0	0	0.80	0.90	0.90	0.90
28	2301	3439	1612	1943	1	0	0	0	0.90	0.90	0.90	0.80
29	2037	2804	1737	1256	0	0	0	0	1.00	1.00	1.00	0.90
30	1025	1331	1436	985	0	0	0	0	0.90	1.00	0.80	0.90
31	2181	3515	1700	1665	0	0	0	0	0.90	0.90	1.00	0.90
32	2376	3493	1818	1865	0	0	0	0	0.90	1.00	0.90	0.80
33	2119	3032	1719	1349	0	0	0	0	0.90	0.80	1.00	0.60
34	2861	4179	1765	1942	0	0	0	0	0.90	1.00	0.90	1.00
35	2084	3317	1787	1453	0	0	0	0	0.90	0.90	0.90	0.90
36	2772	4108	1783	1890	1	1	0	0	0.80	0.90	0.90	0.90
37	2649	4204	1930	2573	0	0	0	0	1.00	0.90	1.00	0.90
38	2282	3349	1840	1976	0	0	0	0	0.90	0.80	1.00	0.60
39	2093	3211	1657	1900	0	0	0	0	0.90	0.90	1.00	1.00
40	2660	3863	1904	2038	0	0	0	0	0.90	1.00	0.90	0.90
41	2156	3032	1676	1642	0	0	0	0	0.80	0.80	0.90	1.00

Tabella1.11S: indici di convergenza delle analisi effettuate usando 10 loci. Gli indici di convergenza sono stati calcolati sulla distribuzione a posteriori totale (Post), sulla distribuzione a priori totale (Pri), sulla likelihood (Lik) e sulla distribuzione a posteriori del numero di cambiamenti demografici (N° C).

	Media ESS				N° ESS < 100				% Geweke non significativi			
Modello	Post	Pri	Lik	N° C	Post	Pri	Lik	N° C	Post	Pri	Lik	N° C
1	1110	1672	817	460	1	1	1	1	0.70	0.80	0.70	0.60
2	1468	2369	981	752	0	0	0	1	0.90	0.90	0.90	0.80
3	1406	2148	942	810	0	0	0	1	0.80	0.80	0.70	0.80
4	976	1553	869	493	1	1	0	2	0.80	0.80	0.90	0.80
5	1040	1524	820	506	4	3	0	4	0.50	0.60	0.70	0.60
6	1850	2992	922	1138	0	0	0	0	0.80	0.70	0.90	0.80
7	1747	2615	1036	880	0	0	0	1	0.90	1.00	1.00	1.00
8	1350	2106	894	732	0	0	0	0	0.90	0.80	1.00	0.80
9	1424	2124	872	786	1	1	0	1	0.90	0.80	1.00	0.80
10	492	612	816	501	1	1	0	0	0.60	0.50	0.70	0.70
11	1203	1721	983	479	0	0	0	1	0.80	0.80	0.90	0.80
12	533	678	790	615	0	0	0	0	0.70	0.70	0.90	0.70
13	614	906	662	602	3	3	0	0	0.60	0.60	0.70	0.90
14	577	813	817	511	1	0	0	0	0.90	0.90	0.90	1.00
15	419	684	575	413	5	3	0	2	0.30	0.40	0.50	0.40
16	1633	2561	1098	1328	0	0	0	0	0.90	0.80	0.80	0.90
17	1463	2382	1024	1261	0	0	0	1	0.80	0.90	1.00	1.00
18	1520	2368	1078	987	0	0	0	1	1.00	0.90	1.00	0.70
19	820	1330	769	612	1	1	0	1	0.60	0.60	0.80	0.60
20	1309	1900	936	756	0	0	0	0	0.90	0.90	0.90	0.80
21	1315	1975	980	896	1	1	0	0	0.90	0.90	1.00	1.00
22	1276	1985	946	803	1	1	0	1	0.80	0.90	1.00	1.00
23	1177	1682	955	718	1	1	0	3	0.90	1.00	0.90	0.80
24	1708	3088	936	882	1	0	0	1	0.90	0.90	0.80	0.80
25	1150	1736	908	717	0	0	0	0	1.00	1.00	0.90	0.80
26	1241	1876	938	662	0	0	0	1	0.90	0.70	1.00	0.80
27	1223	1755	939	760	1	1	0	0	0.80	1.00	0.90	0.80
28	1137	1793	954	579	1	1	0	1	0.80	0.80	0.90	1.00
29	1068	1701	912	693	0	0	0	0	0.80	0.80	1.00	0.70
30	500	590	806	575	0	0	0	0	0.90	0.90	0.80	1.00
31	429	708	660	330	2	1	0	0	0.70	0.90	0.80	0.60
32	1647	2708	1036	1199	0	0	0	0	0.90	1.00	1.00	0.90
33	1021	1469	882	596	0	0	0	0	1.00	1.00	0.90	0.90
34	1465	2242	964	965	0	0	0	1	0.90	1.00	0.90	0.90
35	1528	2487	918	1032	0	0	0	1	0.80	0.80	0.90	0.70
36	1413	2549	920	919	1	1	0	1	0.80	1.00	0.70	0.90
37	1101	1635	917	581	0	0	0	2	0.60	0.60	0.70	0.70
38	1269	1920	995	1057	1	0	0	0	0.60	0.70	0.80	0.70
39	1081	1817	799	809	2	1	0	3	0.70	0.70	0.90	0.80
40	1690	2601	1051	1146	0	0	0	0	1.00	0.90	1.00	1.00
41	1272	2001	982	906	0	0	0	0	0.70	1.00	0.80	1.00

APPLICAZIONE NUMERO DUE

L'IMPORTANZA DEL DNA ANTICO NEL RICOSTRUIRE LA STORIA DEMOGRAFICA DELLE POPOLAZIONI

2.1 INTRODUZIONE

La variabilità genetica osservabile in un campione di sequenze di DNA moderne contiene numerose informazioni sulla storia demografica della popolazione studiata. Negli ultimi anni, riuscire a studiare questo processo e stimarne alcuni parametri chiave è diventato uno degli obiettivi principali in genetica di popolazione con numerose applicazioni in antropologia, biologia della conservazione, epidemiologia e virologia (Drummond et al. 2005; Fagundes et al. 2007; Magiorkinis et al. 2009; Stiller et al. 2010). In particolare, la stima della dimensione effettiva di una popolazione, e come questa sia cambiata nel corso del tempo, risulta essere estremamente informativa riguardo la storia demografica ed evolutiva di una popolazione. Conoscere la demografia passata di una popolazione ha notevoli ricadute pratiche soprattutto in genetica della conservazione, per identificare quelle specie che richiedano piani di protezione, nel caso di specie in declino demografico, o piani di controllo per quelle specie in rapida espansione demografica come le specie invasive.

Le informazioni contenute in campioni di DNA moderno sono comunemente usate per ricostruire eventi demografici avvenuti nel passato (Fagundes et al. 2007; Neuenschwander et al. 2008), ma negli ultimi 25 anni, i progressi effettuati nella genotipizzazione del DNA degradato hanno permesso a molti laboratori di studiare la variabilità genetica anche in campioni antichi, di età compresa dalle decine alle decine di migliaia di anni nel passato. Con il termine "DNA antico" (aDNA) si intende il DNA proveniente dal materiale biologico di un individuo non più in vita. Questo materiale include resti sub fossili (di solito ossa e denti), resti archeologici, coproliti, mummie, reperti congelati naturalmente (ad esempio nel permafrost), sedimenti, erbari e reperti museali. Questo tipo di dato ha di fatto aperto le porte all'analisi genetica delle popolazioni del passato e si è rivelato particolarmente utile in genetica di conservazione per la gestione di specie in pericolo di estinzione a causa dell'intervento antropico soprattutto negli ultimi 100/200 anni (vedi esempi in Wandeler et al. 2007 e Leonard 2008). Le popolazioni di piccole dimensioni risentono in maniera più marcata l'effetto della deriva genetica casuale e perdono rapidamente diversità

genetica. Questo fenomeno rende più difficile determinare le caratteristiche genetiche della popolazione nel passato utilizzando solo DNA moderno. Il aDNA permette di misurare direttamente la diversità genetica nel passato e quindi di ottenere stime della dimensione effettiva passata, livelli di flusso genico, e parentela tra le popolazioni, prima della perturbazione demografica. Alcuni esempi dell'utilizzo del DNA antico in genetica di conservazione includono: la stima della dimensione effettiva del lupo grigio in Nord America prima e dopo la sua completa estirpazione in alcune aree geografiche (Leonard et al. 2005); la stima del livello di introgressione tra le popolazioni di trote (*Salmo trutta*) naturali ed allevate utilizzando reperti museali delle popolazioni di trota prima delle pratiche di allevamento (Hansen 2002) e l'identificazione delle popolazioni più adatte per attuare programmi di reintroduzione di *Oxyura leucocephala* in Spagna (Muñoz-Fuentes et al. 2005), del grizzly di Yellowstone in Nord America (Miller e Waits 2003) e di *Gypaetus barbatus* in Europa (Godoy et al. 2004). Allo stesso modo, verificare il livello di variabilità genetica nel passato e nel presente, può essere utile per identificare espansioni demografiche. Questo tipo di fenomeno, come anche le reintroduzioni e le invasioni, può talvolta mettere in contatto specie allopatriche che non hanno sviluppato barriere riproduttive e facilitarne la loro ibridazione. Gli effetti dell'ibridazione sono difficilmente prevedibili ma possono essere estremamente dannosi per le specie native come documentato per numerosi gruppi tassonomici comprendenti pesci (Perry et al. 2002), uccelli (Mank et al. 2004; Munoz-Fuentes et al. 2007), mammiferi (Leonard e Wayne 2008), e piante (Saltonstall 2002).

Il recente sviluppo della teoria Coalescente (Kingman 1982a; Kingman 1982b) e la sua successiva estensione nel Coalescente Seriale (Rodrigo e Felsenstein 1999), hanno permesso lo studio della storia demografica di una popolazione integrando dati molecolari moderni ed antichi. Simulazioni coalescenti sono state utilizzate per confrontare la performance di un campione di DNA mitocondriale moderno e un campione formato da DNA moderno e antico nell'identificare una riduzione demografica avvenuta 2 000 generazioni nel passato (Ramakrishnan et al. 2005). Il campione genetico comprendente aDNA si è dimostrato migliore nell'identificare la riduzione demografica rispetto al solo campione moderno e il potere generale è risultato essere influenzato dall'intensità della riduzione demografica e dalla dimensione effettiva dopo la riduzione demografica. Sebbene questo studio fornisca prime indicazioni sull'effetto dell'inclusione di aDNA nelle analisi, le sue conclusioni sono limitate a un singolo marcatore genetico e l'effetto di usare un marcatore diverso non è prevedibile. Inoltre il confronto dei due campioni è basato sulle distribuzioni di due indici di variabilità genetica (Eterozigosità e D di Tajima) che non riassumono tutta l'informazione presente nei dati, ma solo una parte, per cui i risultati ottenuti sembrano essere indicativi di un trend generale e non generalizzabili ad altri modelli demografici. Inoltre, le

performance sono state misurate solamente secondo uno schema di campionamento contenente DNA moderno e uno schema contenente DNA moderno e DNA antico campionato subito prima della riduzione demografica, perciò l'effetto di altri tipi di campionamento (ad esempio avere più campioni temporali) rimane sconosciuto. L'inclusione di aDNA sembra essere importante anche nel processo di stima di parametri demografici. Chan et al. (2006) hanno dimostrato che includere campioni antichi nell'analisi rende possibile la stima dell'intensità della riduzione demografica e della dimensione effettiva antica di *Ctenomys sociabilis* in modo più preciso che se stimato con solo DNA moderno.

Sebbene il aDNA sia uno strumento molto utile per ottenere informazioni genetiche da individui non più presenti, presenta però alcuni svantaggi. La possibilità di includere nelle analisi un campione antico dipende soprattutto dalla disponibilità in natura o in museo di reperti appartenenti alla specie d'interesse, per cui in alcuni casi è semplicemente impossibile il suo utilizzo. Inoltre, il reperto biologico deve essere distrutto per poterne estrarne il DNA, impedendo l'analisi di tutti quei reperti per cui la disponibilità nei musei è molto scarsa o unica. La successiva tipizzazione genetica richiede delle misure di controllo non standard per evitare il rischio di contaminazione con DNA moderno e una fase di validazione del dato genetico ottenuto piuttosto lunga. L'insieme di questi fattori fa si che la produzione in laboratorio del dato genetico antico sia dalle cinque alle dieci volte maggiore rispetto al DNA moderno, senza considerare il costo per la raccolta dei campioni in natura o dalle collezioni museali. Studiare le condizioni in cui uno studio di genetica di popolazione trae vantaggio dall'inclusione di campioni antichi al suo interno, risulta un problema di notevole importanza per non sprecare risorse economiche e campioni di valore storico. Quantificare quanto la combinazione di DNA moderno ed antico aumenti il potere di identificare una riduzione o un aumento demografico è ancora poco studiata ed inoltre non esistono informazioni in letteratura riguardanti l'effetto di diversi schemi di campionamento temporale antico sull'inferenza.

L'obiettivo principale di questo studio è di valutare il contributo dell'utilizzo del DNA antico nell'identificare una riduzione demografica. Simulazioni coalescenti seriali sono state utilizzate per generare sequenze di DNA da una popolazione in declino demografico e combinate secondo diversi schemi di campionamento ognuno dei quali caratterizzato da una diversa quantità e distribuzione temporale del campione antico. Il metodo dell'Extended Bayesian Skyline Plot (Heled e Drummond 2008) è stato successivamente utilizzato per valutare l'abilità di ogni schema di campionamento nel ricostruire una riduzione demografica tenendo in considerazione il maggior costo di produzione del DNA antico. I risultati ottenuti indicano che non solo includere DNA antico nell'analisi sembra contribuire in modo significativo alla precisa ricostruzione del modello evolutivo

ma anche la distribuzione temporale delle sequenze antiche sembra giocare un ruolo chiave nel processo di inferenza.

2.2 MATERIALI E METODI

Lo studio di simulazione, si compone di tre passaggi principali: i) la simulazione dei dati genetici secondo diverse condizioni demografiche; ii) la ricostruzione della funzione demografica (relazione tra dimensione effettiva della popolazione e il tempo in generazioni) a partire dai dati di variabilità molecolare e iii) il confronto fra la dinamica simulata e ricostruita, la quantificazione dell'errore nella ricostruzione e il confronto dei diversi schemi di campionamento in accordo con l'accuratezza e la precisione della ricostruzione.

Passaggio 1: simulazione dei dati genetici

100 dataset di DNA sono stati simulati con il software SerialSimcoal (Anderson et al. 2005) per ogni specifico scenario di simulazione. Uno scenario di simulazione è stato definito come la combinazione di un modello demografico (con i parametri demografici associati), dal tipo e numero di marcatori genetici e da uno schema di campionamento temporale. Il modello demografico è stato creato per riprodurre una riduzione demografica avvenuta in tempi recenti, come documentato per numerose specie animali soprattutto a causa dell'intervento antropico. Il modello prevede una popolazione di dimensione stabile che subisce una riduzione istantanea, che ne fa decrescere la dimensione di un certo numero di volte, e poi rimane stabile fino al presente. Per molte specie animali, il tempo di generazione varia dai 2 ai 4 anni (Mech e Seal 1987; Nichols et al. 2001; Chan et al. 2006), perciò il tempo del cambiamento demografico (T_C) è stato fissato a 50 generazioni nel passato (100/200 anni). Gli altri parametri del modello sono la dimensione effettiva moderna (N_0), la dimensione effettiva pre-riduzione (N_A) e l'intensità del cambiamento demografico ($I_C=N_A/N_0$). Le dimensioni effettive moderne sono state scelte in modo da avere in media al momento della riduzione (50 generazioni nel passato) il 50%, 90% o 99% delle linee genealogiche moderne. Seguendo quanto derivato da Tavaré (1984), è possibile calcolare il valore atteso del numero di linee genealogiche presenti a un certo momento nel passato:

$$E(A_t|A_0 = i) = \sum_{l=1}^{i} p_l^0\,(t)\,\frac{(2l-1)i_{[l]}}{i_{(l)}}$$

$$p_l^0(t) = exp\{-k(k-1)t/2\}$$

$$i_{(l)} = i(i+1) \ldots (i+l-1), \;\; l \geq 1; \;\; i_{(0)} = 1$$

$$i_{[l]} = i(i-1) \ldots (i-l+1), \;\; l \geq 1; \;\; i_{[0]} = 1$$

dove *i* è il numero di linee genealogiche al tempo 0 (presente) e *t* è il tempo dell'evento demografico scalato per la dimensione effettiva della popolazione. L'obiettivo è stato raggiunto impostando le dimensioni effettive a 150, 750 e 1500 individui aploidi. Sono state prese in considerazione tre intensità dell'evento demografico: una riduzione debole che porta al dimezzamento della dimensione effettiva antica (I_C=2) e due riduzioni più intense dove la popolazione moderna è 10 o 100 volte più piccola rispetto a prima della riduzione (I_C=10, 100). Una rappresentazione schematica dei modelli e dei loro parametri associati è presente in Figura2.1.

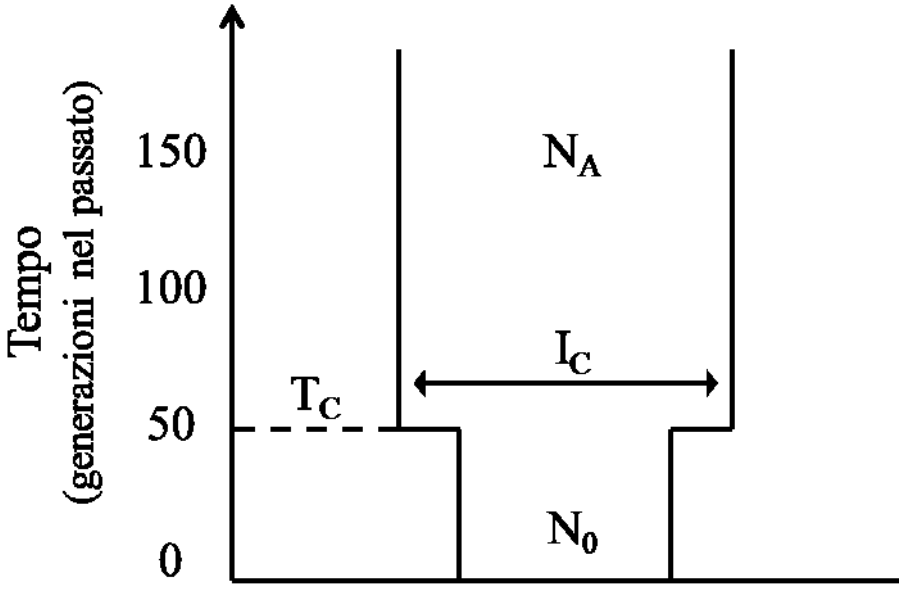

Figura2.7: Illustrazione grafica del modello di riduzione demografico e dei parametri associati. N_A: dimensione effettiva della popolazione prima della riduzione demografica; N_0: dimensione effettiva dopo la riduzione demografica; T_C: tempo della riduzione demografica (fissato a 50 generazioni nel passato); I_C: intensità della riduzione demografica (con lo scorrere del tempo dal presente al passato, IC indica di quante volte la popolazione antica era maggiore di quella moderna).

Una sequenza di 400 paia di basi è stata simulata secondo un modello mutazionale a siti finiti usando un tasso di mutazione di $1x10^{-6}$ mutazioni per sito per generazione (Soares et al. 2009) e un tempo di generazione di 2 anni, in modo da replicare una sequenza proveniente da un locus di DNA mitocondriale, usato comunemente in genetica di conservazione. La dimensione del campione è stata fissata a 80 individui aploidi.

L'ultimo elemento necessario per la definizione di uno scenario di simulazione è lo schema di campionamento temporale. Uno schema di campionamento temporale è una particolare distribuzione di *n* sequenze di DNA nell'intervallo di tempo tra il presente (t=0) e l'età del

campione più antico a disposizione (fissato a t=250) espressa in generazioni. Ad esempio, se n=80, uno schema di campionamento potrebbe essere composto da tutte le sequenze al tempo 0 cioè un campione di sole sequenze moderne oppure potrebbe assegnare 40 sequenze al tempo 0 e 40 sequenze a 100 generazioni nel passato formando un campione composto da DNA moderno ed antico. Analizzare tutte le possibili disposizioni di *n* sequenze in un arco temporale discreto, formato da 251 possibili valori (0,1,2,…,250), non è però trattabile dal punto di vista computazionale ed è perciò stato ristretto a 16 tempi discreti di campionamento temporale tra t=0 e t=250 assegnando ad ogni punto temporale un uguale numero di sequenze. Un sottoinsieme formato da 42 schemi di campionamento è stato in fine analizzato (vedi Tabella 2.1).

Tabella2.1: descrizione degli schemi di campionamento utilizzati. Il gruppo rappresenta il numero di campionamenti temporali. Il sottogruppo indica la posizione dei campionamenti rispetto al tempo della riduzione demografica, con "r" sono indicati i campioni pre-riduzione, con "o" quelli post-riduzione. I tempi da T1 a T16 rappresentano i 16 possibili tempi di campionamento ordinati dal piu antico al più recente.

CODICE	GRUPPO	SOTTOGRUPPO	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16
8	1C	controllo	200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	1C	controllo	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	1C	controllo	50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	1C	1o	40	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	1C	1o	30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	1C	1o	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	1C	1o	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1	1C	1o	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
27	2C	1r1o	200	40	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24	2C	1r1o	200	30	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	2C	1r1o	200	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	2C	1r1o	200	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15	2C	1r1o	200	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
26	2C	1r1o	100	40	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23	2C	1r1o	100	30	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20	2C	1r1o	100	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	2C	1r1o	100	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14	2C	1r1o	100	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
25	2C	1r1o	50	40	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	2C	1r1o	50	30	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	2C	1r1o	50	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16	2C	1r1o	50	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	2C	1r1o	50	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
33	2C	2o	40	30	-	-	-	-	-	-	-	-	-	-	-	-	-	-
32	2C	2o	40	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
30	2C	2o	40	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	2C	2o	40	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
31	2C	2o	30	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
29	2C	2o	30	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	2C	2o	30	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
28	2C	2o	20	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	2C	2o	20	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	2C	2o	10	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
35	4C	3r1o	200	100	50	0	-	-	-	-	-	-	-	-	-	-	-	-
40	4C	2r2o	200	100	40	20	-	-	-	-	-	-	-	-	-	-	-	-
37	4C	2r2o	200	50	40	0	-	-	-	-	-	-	-	-	-	-	-	-
36	4C	2r2o	200	50	20	0	-	-	-	-	-	-	-	-	-	-	-	-
39	4C	1r3o	200	40	20	0	-	-	-	-	-	-	-	-	-	-	-	-
38	4C	1r3o	50	40	20	0	-	-	-	-	-	-	-	-	-	-	-	-
34	4C	4o	40	30	10	0	-	-	-	-	-	-	-	-	-	-	-	-
41	8	3r5o	200	100	50	40	30	20	10	0	-	-	-	-	-	-	-	-
42	16	6r10o	250	200	150	100	75	50	45	40	35	30	25	20	15	10	5	0

Il sottoinsieme è stato scelto arbitrariamente in modo da rappresentare gli schemi di campionamento considerati più importanti: Tra gli otto schemi con un solo campionamento temporale (1C, Tabella2.1), gli schemi 8,7,6 rappresentano un controllo (essendo composti da sequenze di DNA campionate precedentemente all'evento demografico), lo schema 1 contiene solo DNA moderno mentre negli schemi 2,3,4,5 le sequenze sono state prese successivamente alla riduzione demografica ma non al tempo 0; venticinque combinazioni con due campioni temporali formano gruppo 2C (Tabella2.1,codice 9-33), di cui 10 prima dell'evento demografico (sottogruppo 1o, Tabella2.1, codice 19-12,28-33,) e 15 con un campione prima e uno dopo l'evento (sottogruppo 1r10,Table1, codice 13-27); sette schemi con quattro campionamenti temporali distribuiti prima dell'evento (codice 34), uno prima e tre dopo (codice 35), due prima e due dopo (codice 36,37,40), o tre prima e uno dopo (codice 38,39); e infine uno schema con otto (codice 41) e uno con sedici (codice 42) campionamenti lungo tutto l'arco temporale. Seguire quest'approccio è stato necessario per limitare la dimensione totale del problema. Considerando che i) un singolo modello demografico è stato analizzato, ii) il modello demografico è caratterizzato da diverse combinazioni di parametri, iii) sono stati definiti 42 schemi di campionamento, iv) 100 repliche sono state analizzate per ogni scenario simulativo, 37.800 data set genetici sono stati prodotti e analizzati statisticamente.

Passaggio 2: la ricostruzione della funzione demografica

Tutti i metodi definiti nella classe degli "Skyline Plot" prevedono due passaggi distinti e separabili. La genealogia degli individui analizzati deve essere ricostruita a partire dai dati genetici ed include non solo la stima delle relazioni tra gli individui (topologia dell'albero) ma anche i loro tempi di divergenza (età dei nodi). Questo passaggio può essere compiuto utilizzando i metodi filogenetici standard Bayesiani o di massima-verosimiglianza. Una condizione essenziale è che le lunghezze dei rami dell'albero siano proporzionali con il tempo, perciò il tempo deve essere scalato in mutazioni, anni o generazioni. La genealogia, essendo una stima basata su un campione di sequenze di DNA, porta con se un errore, chiamato "errore filogenetico", che può essere di notevole entità quando la genealogia contiene rami interni corti. Inoltre, molti organismi sono caratterizzati da una bassa variabilità genetica intraspecifica che provoca un aumento della varianza stocastica nella lunghezza dei rami. Nonostante questi fattori, se vogliamo ricostruire la storia demografica di una popolazione, non serve che la genealogia sia ben risolta, soprattutto quando le stime sono pesate tra un grande numero di alberi come nel framework Bayesiano (Drummond et al. 2005).

Il secondo passaggio prevede la stima della storia demografica basata sulla genealogia stimata. Una caratteristica molto utile di questa fase è che dipende solamente dal tempo degli eventi di coalescenza e non dal fatto di ricostruire la genealogia esatta delle sequenze del campione (Pybus et al. 2000). Per esempio, osservare eventi di coalescenza molto vicini tra loro è indicativo di una dimensione effettiva piccola, e questo principio può essere sfruttato per stimare la dinamica della dimensione effettiva. Più precisamente, i metodi "Skyline Plot" si basano sulla semplice relazione tra la dimensione effettiva della popolazione e la lunghezza attesa degli intervalli di coalescenza secondo il modello Coalescente: la dimensione effettiva media in ogni intervallo tra due eventi di coalescenza può essere stimata dal prodotto della lunghezza dell'intervallo (γ_i) e $i(i-1)/2$, dove i rappresenta il numero di linee dell'albero presenti nell'intervallo (Figura2.2a). In questo modo è possibile ottenere una stima della dimensione effettiva in ogni intervallo (definito da ogni evento di coalescenza) della genealogia stimata (Figura2.2b) e così ricostruire la storia demografica tramite la dinamica della dimensione effettiva intervallo dopo intervallo. La ricostruzione demografica include una considerevole parte di incertezza dovuta alla natura stocastica del Coalescente. Infatti, ogni genealogia considerata è solo una singola realizzazione casuale di questo processo e questo comporta che, ad esempio, la stima della dimensione effettiva in ogni intervallo di coalescenza incorpori una notevole quantità di errore. L'errore dovuto al Coalescente è inversamente proporzionale al numero di linee genealogiche presenti in ogni intervallo di coalescenza e quindi non è uniforme lungo la genealogia: più ci avviciniamo alla radice dell'albero, più aumenta l'errore nella stima della dimensione effettiva. Ad esempio, l'ultimo intervallo di coalescenza (γ_2 in Figura2.2) viene stimato a partire da sole due linee ed è perciò l'intervallo con più errore associato. Questo fatto diventa perciò di notevole importanza ad esempio quando si sta considerando una popolazione costante dove, in media, l'ultimo intervallo di coalescenza occupa metà della genealogia.

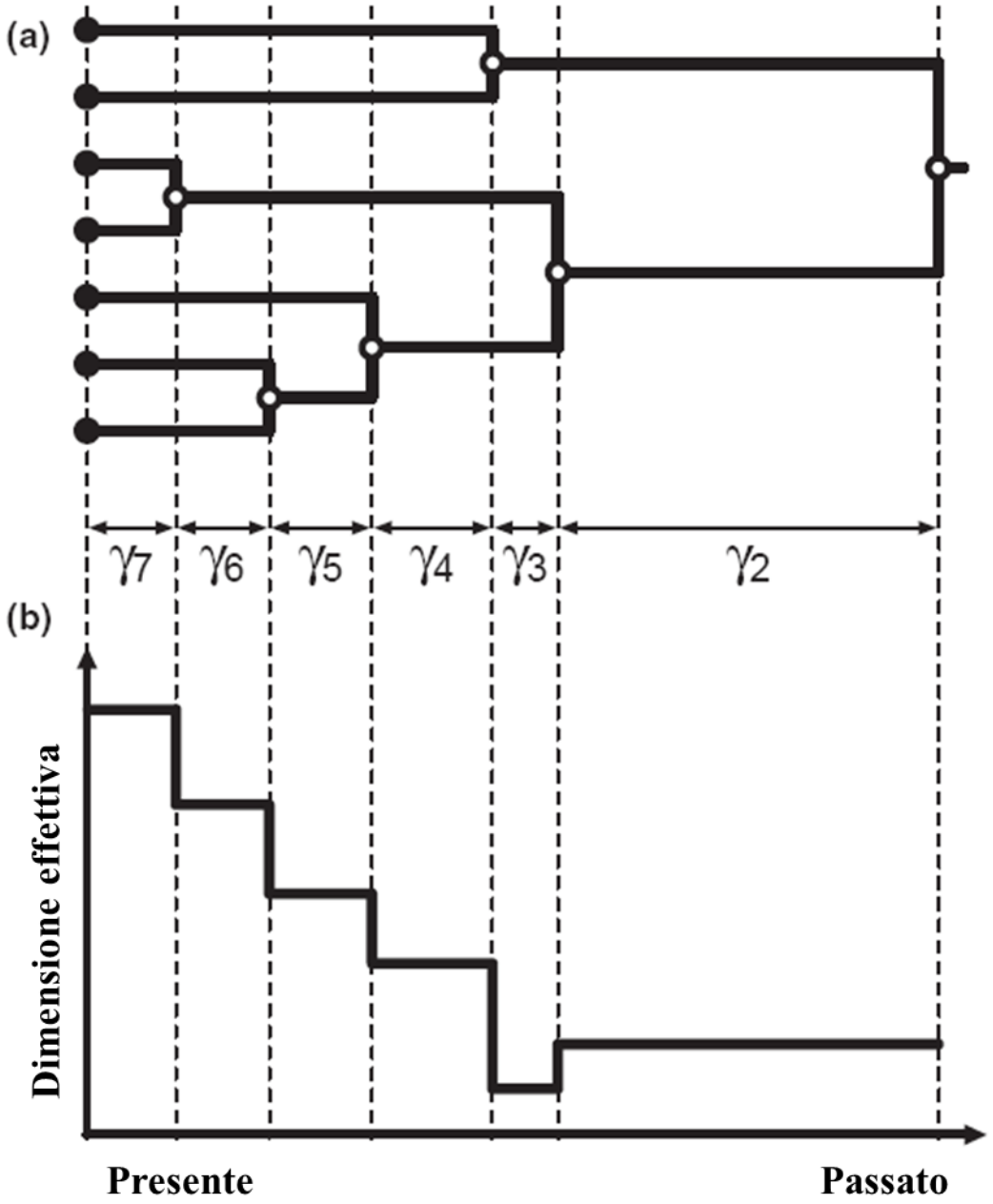

Figura2.8: Stima della storia demografica da una genealogia. (a) Una genealogia stimata a partire dai dati genetici, dove la lunghezza di ogni ramo è proporzionale al tempo. Gli intervalli di coalescenza sono indicati con γ_i. (b) La dimensione effettiva della popolazione stimata in ogni intervallo di coalescenza.

Una volta ottenuta la stima della dimensione effettiva in ogni intervallo, la dinamica della dimensione effettiva nel tempo viene ricostruita in principalmente due modi (vedi review Ho e Shapiro 2011 per i dettagli dei metodi): i) partendo dall'intervallo più recente, la dimensione effettiva viene unita con quella dell'intervallo adiacente assumendo che sia rimasta costante all'interno dell'intervallo, producendo un grafico a "scalini" simile a quello rappresentato in Figura2.2b; ii) la dimensione effettiva non rimane costante all'interno dell'intervallo, ma varia in

maniera lineare tra due intervalli, rappresentando in maniera più realistica come una popolazione aumenta o si riduce rispetto a un cambiamento istantaneo. Nei metodi "Skyline Plot" bayesiani, la dimensione effettiva in ogni intervallo non è descritta da un singolo valore ma bensì da una distribuzione, detta distribuzione a posteriori. Perciò, per ogni intervallo di coalescenza viene utilizzato un indice di tendenza centrale come la media, moda o mediana come stima della dimensione effettiva. Insieme alla stima viene riportato anche l' "High Posterior Density Interval" (HPD, o intervallo di credibilità) che descrive qual è l'intervallo più piccolo che contiene il 90% o il 95% dei valori a maggior frequenza (vedi Figura2.3).

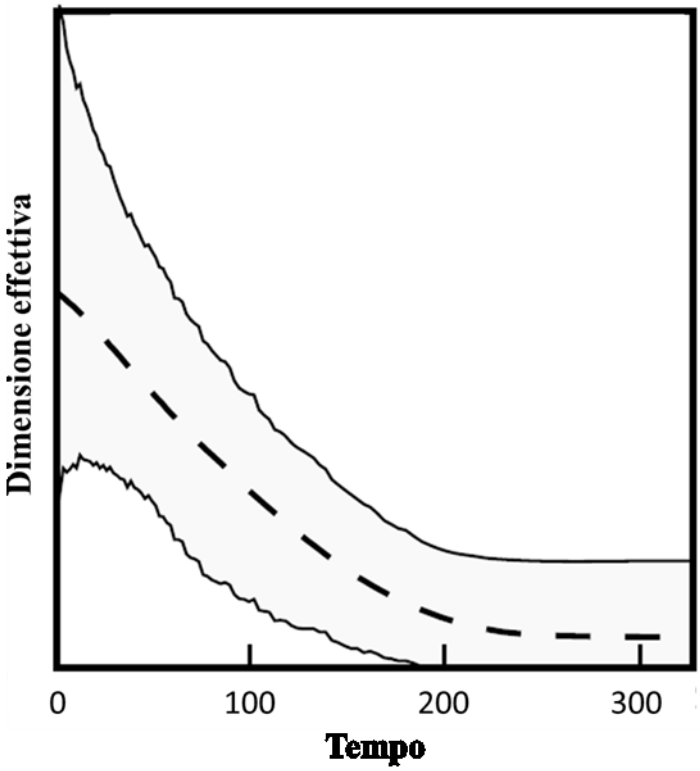

Figura2.9: Rappresentazione grafica di uno skyline plot. La dimensione effettiva (in ordinata) è visualizzata in funzione del tempo (in ascissa). Il tempo è misurato dal presente (0) fino al tempo dell'antenato comune più recente (TMRCA). La dimensione effettiva La linea tratteggiata rappresenta la mediana della ricostruzione mentre in grigio è evidenziato l'HPD.

In questo studio la dinamica della dimensione effettiva nel tempo è stata stimata con il metodo "Extended Bayesian Skyline Plot" implementato nel software BEASTv1.6.1 (Drummond et al. 2002; Heled e Drummond 2008). Questo metodo, basato su un framework bayesiano accoppiato a Monte Carlo Markov Chain (MCMC), permette di stimare la genalogia degli individui simulati a partire dai dati di variabilità molecolare e ricostuire la funzione demografica nel tempo in un singolo passaggio. Inoltre, a differenza degli altri "Skyline Plot", si possono analizzare

contemporaneamente più loci indipendenti per stimare la storia demografica di una popolazione, riducendo in questo modo l'errore associato al Coalescente e, di conseguenza, l'errore nella stima (si riduce l'ampiezza del HPD). Ad ogni locus impiegato nell'analisi è possibile associare un fattore che tenga in considerazione la sua ploidia ed ereditabilità. In questo modo, ad esempio, è possibile tenere in considerazione che la dimensione effettiva di un locus autosomale trasmesso in maniera biparentale è quattro volte maggiore rispetto ad un locus aploide mitocondriale. Inoltre, solo utilizzando l'"Extended Bayesian Skyline Plot" è possibile stimare il numero di cambiamenti demografici tramite "Bayesian Stochastic Variable Selection" (BSVS, Kuo e Mallick 1998). La presenza di molti intervalli di coalescenza corti può portare a un notevole aumento del "rumore di fondo" nella ricostruzione demografica perciò Strimmer e Pybus (2001) proposero di eliminare gli intervalli troppo corti raggruppandoli con i loro vicini in un numero *n* di gruppi specificato a priori. Con il BSVS è possibile selezionare l'*n* maggiormente supportato dai dati senza doverlo necessariamente specificare arbitrariamente a priori. Se un solo gruppo viene selezionato, significa che i dati supportano uno scenario di popolazione costante nel tempo.

Ognuno dei 37 800 dataset genetici è stato convertito in un input file leggibile da BEAST creando un modello con il software BEAUTI (Drummond e Rambaut 2007) e inserendo nel modello i dati molecolari attraverso uno script per l'ambiente per l'analisi statistica R (R Development Core Team 2010) sviluppato *ad hoc*. Lo stesso modello mutazionale utilizzato per la generazione dei dati genetici simulati è stato impiegato nell'analisi (HKY85, tasso di mutazione fissato a $1x10^{-6}$ mutazioni per sito per generazione, frequenza delle basi uguale e rapporto transizioni su trasversioni uguale a uno) in modo che i risultati non risentano dell'incertezza della stima dei parametri mutazionali. Il tasso di mutazione è stato scalato in generazioni (1 generazione = 2 anni) in modo da stimare il tempo in generazioni e la dimensione effettiva della popolazione in numero di individui aploidi. La distribuzione a priori del numero di cambiamenti demografici è stata definita come una distribuzione di Poisson con media ln(2) in modo da favorire nel 50% dei casi uno scenario di popolazione costante e nel restante 50% dei casi almeno un cambiamento demografico. La lunghezza della catena ha previsto 20 milioni di iterazioni con un campionamento dei parametri del modello ogni 10.000 iterazioni ed è stato scartato il primo 10% della lunghezza totale della catena perché non informativo (*burn-in*). I valori degli operatori che regolano il campionamento MCMC sono stati mantenuti invariati rispetto a quelli di default. Alla fine di ogni analisi è stato valutato se la convergenza è stata raggiunta, cioè se l'algoritmo ha campionato dalla distribuzione a posteriori di ogni parametro. A questo scopo sono state calcolate due misure di convergenza per i parametri più importanti del modello: la likelihood, la distribuzione a priori globale (indica il campionamento complessivo da tutte le distribuzioni a priori di tutti i parametri),

la distribuzione a posteriori complessiva (indica il campionamento complessivo da tutte le distribuzioni a posteriori di tutti i parametri) e numero di cambiamenti demografici. Come primo indice è stato calcolato il valore di "Effective sample size" (ESS) che rappresenta il numero di campionamenti indipendenti dalla distribuzione a posteriori stimata. L'algoritmo di campionamento di tipo MCMC, per sua natura, effettua dei campionamenti che sono correlati tra loro per cui il valore di ESS indica la qualità della stima della distribuzione a posteriori. Un ESS minore di 100 è considerato in genere un valore basso e indicativo di problemi durante l'analisi. Come seconda indicazione di convergenza, è stato eseguito il test di Geweke. Questo test si basa sul principio che se la catena ha raggiunto la convergenza, la prima parte e l'ultima parte della catena avranno la stessa media e la loro differenza sarà distribuita in modo normale. Il comando "geweke.diag" disponibile nel package CODA per l'ambiente statistico R (R Development Core Team 2010) è stato utilizzato per condurre il test e ottenere il p-value associato.

Per ogni scenario di simulazione studiato, sono stati riassunti i dati di convergenza dei 100 dataset simulati calcolando per i quattro parametri considerati: la media di ESS, il numero di dataset che hanno un valore di ESS minore di 100 e la percentuale di dataset che hanno mostrato un p-value del Geweke test non significativo per un valore di α=0.05. Le analisi dei dataset che hanno raggiunto la convergenza sono state utilizzate per ricostruire la dinamica demografica della popolazione nel tempo. Ogni ricostruzione effettuata con BEAST può essere immaginata come una distribuzione di funzioni demogafiche campionate via MCMC (Heled e Drummond 2008). La mediana dei valori di N in n punti, dove n corrisponde alla media dei tempi di coalescenza ordinati in tutte le genealogie campionate, è stata utilizzata per costruire la funzione demografica mediana denominata $N'_{50}(t)$. Lo stesso procedimento è stato usato per calcolare la funzione demografica $N'_{2.5}(t)$ e $N'_{97.5}(t)$ (i limiti dell'HPD95%).

Passaggio 3: il confronto tra la dinamica ricostruita e simulata, e il confronto tra gli schemi di campionamento

Ogni funzione demografica ricostruita è stata confrontata con quella utilizzata per generare i dati simulati. Sono stati calcolati tre indici per rappresentare lo scostamento tra le due funzioni:

Errore relativo di ricostruzione

Confrontando la funzione demografica mediana $N'_{50}(t)$ con la funzione demografica simulata N(t) utilizzando una funzione normalizzante, da t=0 alla mediana della distribuzione

dell'altezza della genealogia ricostruita t = τ, è stato possibile definire l'errore relativo di ricostruzione come:

$$err\big(N'(t), N(t)\big) = \sqrt{\int_0^{\tau} \left(\frac{N'_{50}(t) - N(t)}{N(t)}\right)^2 dt}$$

e rappresenta la distanza media relativa tra una stima e il valore conosciuto (Heled e Drummond 2008), utile per quantificare l'accuratezza della ricostruzione.

Dimensione relativa dell'intervallo di credibilità

Lo stesso principio può essere applicato per calcolare la dimensione media relativa dell'intervallo di credibilità (o HPD) al 95% :

$$dim\big(N'(t), N(t)\big) = \sqrt{\int_0^{\tau} \left(\frac{N'_{97.5}(t) - N'_{2.5}(t)}{N(t)}\right)^2 dt}$$

(Heled e Drummond 2008), utile per quantificare la precisione della ricostruzione.

Coverage della funzione ricostruita

Il coverage dello stimatore (mediana) della funzione ricostruita misura la percentuale di volte in cui il valore della funzione demografica simulata cade nell'intervallo di credibilità al 95% stimato, ed è definito come:

$$cov\big(N'(t), N(t)\big) = \frac{1}{t}\int_0^{\tau} I\big(N'_{2.5}(t) \le N(t) \le N'_{97.5}(t)\big)dt$$

dove *I* è la funzione indicatore (Heled e Drummond 2008).

Questi indici descrivono le proprietà della funzione demografica ricostruita nel suo insieme, ma non sono informativi dello scostamento in regioni locali della ricostruzione demografica. Per questo, la funzione demografica mediana ($N'_{50}(t)$) è stata utilizzata per stimare in ogni simulazione i) la dimensione effettiva pre-cambiamento demografico (N'_0), definita come la media di $N'_{50}(t)$ da t=11 a t=20 (le prime 10 generazioni sono state scartate perché generalmente affette da distorsioni casuali della dimensione effettiva stimata); ii) la dimensione effettiva appena precedente il

cambiamento demografico (N'_{PRE}), definita come la media di $N'_{50}(t)$ da t=106 a t=155; iii) la dimensione effettiva antica (N'_A), definita come la media di $N'_{50}(t)$ negli ultimi due intervalli di coalescenza. Queste tre stime sono state utilizzate per calcolare due stimatori dell'intensità della riduzione demografica:

$$I1 = \frac{N'_{PRE}}{N'_0}, I2 = \frac{N'_A}{N'_0}$$

Inoltre, la funzione demografica mediana N'50(t) è stata utilizzata per ottenere una stima del tempo in cui è avvenuto il cambiamento demografico (T_C). Questo parametro non può essere stimato in maniera esplicita da BEAST perché non fa parte del modello "skyline plot", però può essere stimato indirettamente dalla funzione demografica ricostruita. In particolare si tratta di identificare se, e a quale tempo, la funzione N'50 abbia un punto di flesso. L'approccio analitico comunemente utilizzato consiste nel trovare il punto *x* in cui la derivata seconda della funzione sia uguale a 0 e successivamente verificare se la derivata seconda cambi di segno intorno ad *x*. Tuttavia, alcune analisi preliminari effettuate hanno mostrato che questo sistema analitico risente della presenza di piccole oscillazioni della dimensione effettiva nella regione vicina a t=0 (tipiche delle funzioni demografiche ricostruite), pregiudicandone l'applicabilità. Un metodo alternativo è stato perciò utilizzato. In una prima fase, l'intensità del cambiamento demografico (*I2*) è stata utilizzata come filtro: le funzioni con *I2* < 10% non sono state considerate rappresentare un cambiamento demografico e di conseguenza prive di T_C. Nella seconda fase, in ognuna delle funzioni rimaste, è stata calcolata la derivata prima (*m*) nell'intervallo chiuso [0,TMRCA]. La derivata prima rappresenta il coefficiente angolare della retta tangente alla funzione demografica, ed è perciò informativa sul momento del cambiamento demografico (Figura2.4, b). Nel caso in cui la funzione demografica rappresenti una riduzione (Figura2.4, a), *m* è caratterizzato da un picco positivo sovrapposto al momento del cambiamento demografico (Figura2.4, b). Il tempo del cambiamento demografico (T_c) è stato poi definito come il tempo associato al massimo assoluto della funzione di *m* nell'intervallo delimitato dal momento in cui *m* supera ($T_{c\ inf}$) e ritorna inferiore ($T_{c\ sup}$) a un certo valore soglia. Lo scopo di questo passaggio è identificare rapidi aumenti o riduzioni delle dimensioni effettive nel tempo, perciò il valore di soglia per *m* è stato definito come quello associato alla retta che rappresenta un cambiamento lineare di N'_{50} dal passato al presente. L'intervallo chiuso t[10,20] è stato scelto per rappresentare il presente (sono state scartate le prime 10 generazioni perché soggette ad oscillazioni di N'_{50}) mentre l'intervallo $t[t_{TMRCAmin}-20,t_{TMRCAmin}]$ per il passato. Il parametro $t_{TMRCAmin}$ è il valore di TMRCA minimo all'interno dello scenario simulativo di appartenenza della funzione demografica studiata. Si è scelto di utilizzare $t_{TMRCAmin}$

invece di t_{TMRCA} per avere un valore di TMRCA condiviso tra le funzioni demografiche per ogni scenario simulativo ed evitare di studiare valori di t non previsti in alcune ricostruzioni. Il numero di volte in cui, tra le 100 repliche, è stato possibile identificare una riduzione demografica in questo modo (P_{RID}), rappresenta una stima frequentista della probabilità di identificare un evento demografico ed è stata perciò calcolata per ogni schema di campionamento in ogni scenario simulativo.

Le funzioni demografiche $N'_{50}(t)$, $N'_{2.5}(t)$ e $N'_{97.5}(t)$ sono state ottenute dal'output di Beast interpolando i valori di dimensione effettiva e il tempo espresso in generazioni con una funzione spline, utilizzando il pacchetto "akima" per l'ambiente statistico R (R Development Core Team 2010).

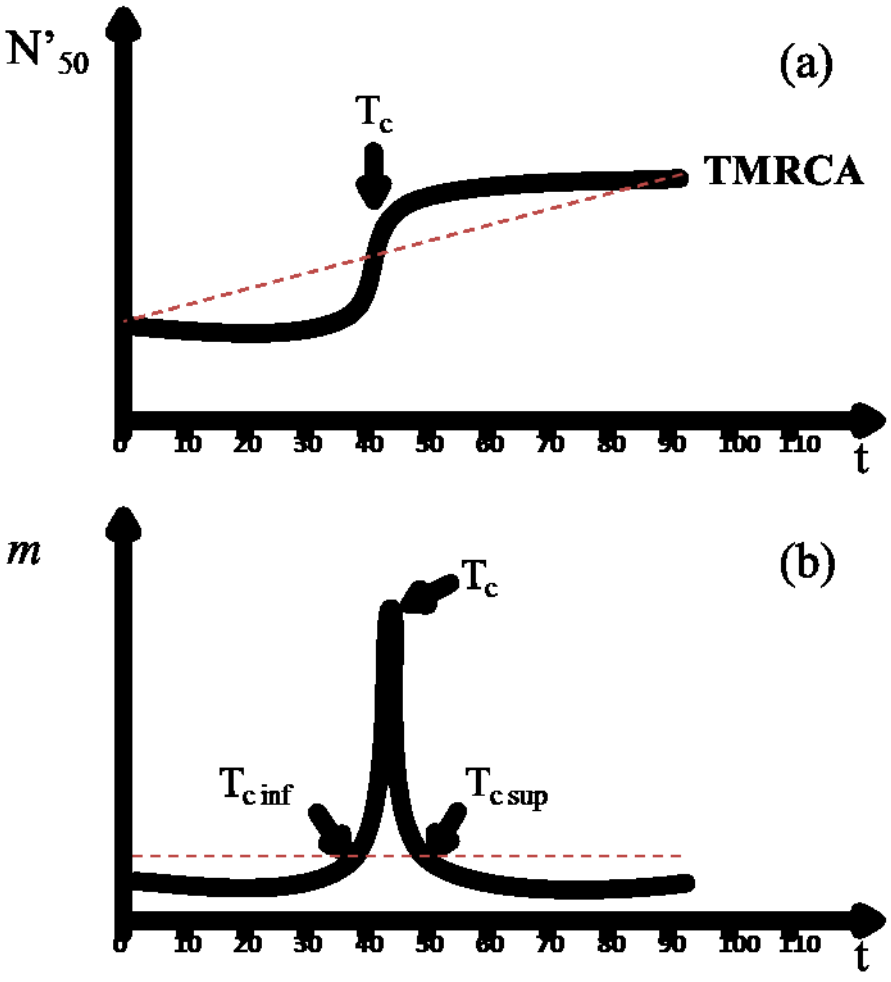

Figura2.10: (a) funzione demografica mediana tipica di una popolazione che ha subito una riduzione demografica (linea nera continua) e la retta di declino lineare tra il TMRCA e il presente (linea rossa tratteggiata); (b) andamento della derivata prima di N'_{50} (linea nera continua) e della retta di declino (linea rossa tratteggiata) che funge da soglia. Il tempo dalla riduzione (T_C) è stato calcolato come il valore massimo nell'intervallo tra il tempo in cui m supera la soglia da sinistra (Tc inf) e da destra (Tc sup).

In ogni scenario simulativo, le distribuzioni degli indici descritti in precedenza (le 100 repliche) sono state caratterizzate calcolandone la mediane la deviazione standard. Inoltre per i parametri N_0, N_{PRE}, N_a, I_1, I_2 è stato calcolato il BIAS e RMSE (Radice dello scarto quadratico medio) dello stimatore rispetto al valore con cui sono stati simulati i dati, definiti come:

$$BIAS = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\theta}_i - \theta}{\theta}$$

$$RMSE = \frac{1}{\theta}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta\right)^2}$$

dove $\hat{\theta}_i$ è lo stimatore del parametro θ e n è il numero di repliche. Questi indici rappresentano lo scostamento relativo dello stimatore dalla stima esatta di un parametro d'interesse. Un BIAS positivo indica una sovrastima del parametro mentre negativo indica una sottostima.

Infine, per ogni scenario simulativo è stata calcolata la probabilità di rifiutare un modello demografico di popolazione costante (P_C) verificando la percentuale di repliche in cui la distribuzione a posteriori del numero di cambiamenti demografici non conteneva lo 0 (un valore di 0 significa popolazione costante)(Heled e Drummond 2008).

In ogni scenario simulativo, è stato utilizzato un modello di regressione logistica multinomiale per descrivere la relazione tra gli schemi di campionamento (variabile risposta) e 8 indici rappresentanti la funzione demografica (*err*, *dim*, P_C, P_{RID}, T_c, N'_0, N'_{PRE} e N'_{TMRCA}, variabili esplicative). Gli indici *cov*, *I1*, *I2* sono stati esclusi dalla stima a causa della totale correlazione con le altre variabili e perciò non avrebbero aggiunto informazione al processo inferenziale. In seguito, per ogni schema di campionamento è stato possibile calcolare la probabilità di generare la combinazione degli 8 indici coincidenti con la funzione demografica osservata (*err=0*, *dim=0*, P_C=1, P_{RID}=1, T_c=50, N'_0=N_0, N'_{PRE}=N_A, N'_{TMRCA}=N_A), cioè la probabilità di ricostruire la funzione demografica osservata (P_{SS}). P_{SS} somma ad 1 all'interno di ogni scenario simulativo, e permette perciò il confronto diretto di ogni coppia di schemi di campionamento all'interno di esso. Per questo motivo, il valore assoluto di P_{SS} non è da ritenersi particolarmente informativo in quanto dipende dal numero di schemi di campionamento confrontati contemporaneamente (ad esempio la P_{SS} dello schema 1 può non essere la stessa se si studiano 42 o 45 schemi di campionamento) mentre il confronto relativo della P_{SS} appartenente a diversi schemi rimane costante ed è utile per comparare

le loro performance. Una classifica degli schemi di campionamento è stata stilata in accordo con i valori di P_{SS} in ogni scenario simulativo.

2.3 RISULTATI

In Figura2.5 sono stati riassunti graficamente i tre indici principali utili a descrivere l'abilità di uno schema di campionamento nel ricostruire la funzione demografica reale in ognuno dei nove scenari simulativi: la probabilità relativa associata ad ogni schema (P_{SS}), la media della stima del tempo della riduzione demografica (T_C medio) e la probabilità di rifiutare l'ipotesi di popolazione costante (P_C). Nel confronto sono stati inseriti tre schemi di controllo (nr 6,7,8) per avere un punto di riferimento rispetto alla non capacità di ricostruire la funzione demografica corretta.

Schemi di campionamento a un campione temporale (gruppo 1C)

Lo schema composto totalmente da DNA moderno (1) è caratterizzato da PSS prossime a zero in quasi tutti gli scenari simulativi analizzati indicando basse performance quando si utilizza solo DNA moderno rispetto a schemi con DNA antico. Anche nei casi in cui P_{SS} per lo schema 1 sia maggiore degli altri schemi ad un campionamento (Figura2.5, gruppo 1C, N_0 750 - 1500, I_C 10) e degli schemi di controllo, non riesce mai a raggiungere valori comparabili con gli schemi migliori ($P_{SS}(1)=0.019$, $P_{SS}(16)=0.112$). Maggiore è l'intensità della riduzione, maggiore è la capacità dello schema 1 di identificare l'evento demografico. Quando l'intensità della riduzione è bassa ($I_C=2$) lo schema 1 non è in grado di rifiutare l'ipotesi di popolazione costante per nessuna delle dimensioni effettive considerate ($P_C=0\%$, $P_{RID}=0$-9%, Figura2.5c $I_C=2$, $N_0=150$, 750, 1500 e Tabella2.2-10). Anche per intensità $I_C=10$, la probabilità P_C assume valori molto bassi ($P_C=0$ - 5%, N_0 150 - 1500) mentre P_C, basata sulla funzione ricostruita, raggiunge valori da 0 al 79% (Tabella2.2-10), suggerendo che sebbene BEAST spesso ricostruisca una riduzione demografica, lo scenario di popolazione costante non può essere escluso utilizzando solo DNA moderno. Come atteso, ad intensità di riduzione massime ($I_C=100$), la P_C dello schema 1 aumenta in maniera proporzionale alla dimensione effettiva residua dopo la riduzione raggiungendo una $P_C=20\%$ per $N_0=150$, fino al 100% nel caso di $N_0=1500$ (Figura2.5c, Tabella2.2-10). Anche in questi casi però, allo schema 1 viene attribuita una bassa P_{SS} dovuta a un bias nella stima del tempo dell'evento demografico, che mediamente, risulta essere tra le 183.3 e le 285.7 generazioni, cioè dalle 3.6 alle 5.7 volte il valore vero (Figura2.5b, Tabella2.2-10, I_C 100). Gli altri schemi appartenenti al gruppo sono caratterizzati da avere un comportamento del tutto simile allo schema 1 in tutti gli scenari studiati, indicando che

campionare le sequenze di DNA progressivamente più vicino al tempo della riduzione demografica non influenza i risultati.

Schemi di campionamento a due campioni temporali (gruppo 2C)

Il gruppo composto da due campionamenti temporali contiene schemi ad alta P_{SS} in tutti gli scenari simulativi analizzati (Figura2.5a). In sei delle nove combinazioni di parametri, ripartire il campionamento in due parti è sufficiente per raggiungere il massimo delle performance, ottenendo valori di P_{SS} maggiori rispetto ad un campionamento più distribuito (4C,8C e 16C) o concentrato in un punto temporale (Figura2.5a, N_0 750 - 1500, I_C 2-100). Negli altri casi invece, il campionamento distribuito viene selezionato come strategia da preferire, ma con una probabilità solo leggermente superiore rispetto agli schemi 2C (Figura2.5a, $N_0$150, I_C 2 - 100). L'analisi di P_C indica che, più intenso è stato l'evento demografico, più è facile identificarlo usando schemi del gruppo 2C (Figura2.5c). Nel caso in cui l'intensità sia bassa, non si nota un guadagno prestazionale nell'identificare una riduzione rispetto al controllo (Figura2.5c, I_C2), mentre nel caso di I_C=10, alcuni schemi del gruppo 2C raggiungono P_C maggiori sia degli schemi di controllo sia degli schemi ad un campione temporale (Figura2.5c, I_C10, schema 13, P_C =17%). L'uso di questo gruppo di schemi si dimostra particolarmente utile nello stimare i parametri demografici associati ad una riduzione intensa soprattutto quando la dimensione effettiva post-riduzione è piccola, come nel caso di N_0=150. In questo caso, non solo la probabilità di identificare la riduzione è maggiore rispetto agli schemi 1C (P_C 1C=20 - 26%, P_C 2C=75 - 100%), ma anche il tempo della riduzione risulta stimato con maggior precisione (T_C 1C ~ 190, T_C 2C schema 13 = 49.2, Figura2.5b,c, N_0=150, I_C=100).

All'interno del gruppo 2C, il sottogruppo comprendente gli schemi con due campionamenti temporali successivi alla riduzione demografica (sottogruppo 2o), ha mediamente una P_{SS} più bassa rispetto agli schemi con i dati genetici presi prima e dopo la riduzione (sottogruppo 1r1o) (Figura2.5a). Questa differenza di P_{SS} è principalmente dovuta alla stima del tempo della riduzione demografica. Infatti, quando l'intensità dell'evento è massima (I_C=100), ed è perciò più facilmente identificabile, il sottogruppo 2o tende sovrastimare il T_C (circa 4.8 volte nel caso di N_0=1500), e produrre stime della intensità della riduzione demografica più lontane dai valori simulati (Tabella2.2-10). Lo stesso andamento si osserva in misura minore per I_C=10 (N_0 1500) dove T_C tende ad essere stimato mediamente a 76.3 generazioni. Il confronto tra il Relative Recovery Error

(RRE) e il Relative Credible Interval (RCI) all'interno del gruppo 2C, indica che avere un campione prima e dopo l'evento demografico (1r1o) produce stime più precise ed accurate della strategia 2o (Figura2.6). In tutte le combinazioni di parametri demografici, gli schemi 1r1o (in blu) tendono a minimizzare i due indici rispetto a 2o. La differenza tra i due sottogruppi si nota in particolar modo quando la riduzione demografica produce una dimensione effettiva di 150 individui. In questo caso, avere un campione proveniente dalla popolazione prima della riduzione aumenta l'accuratezza della ricostruzione demografica piuttosto della precisione nella stima(Figura2.6, N_0 150). Quando le dimensioni della popolazione post-riduzione sono più grandi (N_0 750 - 1500), il sottogruppo 1r1o risulta essere ancora il migliore ma con una differenza meno marcata rispetto a 2o (Figura2.6, N_0 750 - 1500, I_C=2 - 100). Come atteso, le inferenze prodotte da uno schema di tipo 2o, essendo basate su campioni dalla popolazione post-riduzione, risentono del livello di variabilità genetica nella popolazione in quel momento temporale producendo quindi stime meno accurate se il livello è basso, ma stime più accurate se viene conservata variabilità genetica dopo la riduzione.

All'interno del sottogruppo 1r1o, gli schemi aventi un campione pre-riduzione demografica a 200 generazioni nel passato (27,24,21,18,15), ottengono P_{SS} molto basse, del tutto simili agli schemi di controllo (Figura2.5a). Sebbene questi schemi, a differenza del controllo, riescano a identificare la riduzione demografica a una frequenza simile agli schemi migliori (vedi P_C in Figura2.5c, N_0 750 - 1500, IC10 - 100 e Tabella2.2-10), la stima del T_C sembra essere distorta verso valori superiori al reale, e coincidenti con il tempo del campionamento pre-riduzione demografica (Figura2.5b). Spostando il tempo del campionamento pre-riduzione a 100 generazioni (schemi 26,23,20,17,14), si nota un generale aumento di P_{SS} in tutti gli scenari analizzati, dovuto principalmente alla stima di T_C che tende maggiormente verso il valore vero. Questa stima però sembra essere ancora influenzata dal tempo del campionamento pre-riduzione, assumendo spesso valori tendenti a 100 generazioni, soprattutto negli scenari ad alta intesità (I_C=100) dove la riduzione è più facilmente identificabile. Muovendo il campionamento pre-riduzione ad un valore coincidente con la riduzione demografica si osservano le migliori performance. Gli schemi 25,22,19,16,13 ottengono valori di P_{SS} e P_C tra i più alti del gruppo soprattutto con gli schemi 16 e 13. Questi due schemi, negli scenari con $I_C > 2$, riescono a raggiungere i valori di P_C più alti tra tutti gli schemi confrontati e soprattutto producono una stima di T_C molto vicina al valore simulato (Figura2.5b). In questa serie di schemi, spostare il tempo del campionamento post-riduzione, tenendone invariato il pre-riduzione, ha l'effetto di peggiorare le performance sia nella P_C sia nella stima del T_C (vedi ad esempio lo schema 13 e lo schema 25 in N_0 150 - 1500 I_C 10, N_0 150 - 1500 I_C 100).

Schemi di campionamento a quattro campioni temporali (4C) e a campionamento temporale altamente distribuito (8C, 16C)

Il gruppo composto dagli schemi aventi quattro punti di campionamento temporale mostra buone performance nella ricostruzione della dinamica demografica simulata, ottenendo valori di P_{SS} superiori agli schemi a singolo campionamento in tutti gli scenari simulativi (Figura2.5a). Inoltre, a questo gruppo viene attribuito il massimo valore di probabilità in tutti i casi con N_0=150 (Tabella2.2-10) ma senza evidenziare una grossa differenza dai valori di P_{SS} ottenute con campionamento di tipo 2C. Negli scenari simulativi con N_0=750 - 1500, questa tipologia di campionamento ottiene valori di P_{SS} superati solo dal gruppo di schemi 2C e in un caso da 8C (N_0 750, I_C 10). Questo risultato suggerisce ci sia un vantaggio nel distribuire il campione di sequenze di DNA in quattro punti temporali rispetto a una strategia a singolo campionamento (1C), ma che le sue performance non siano maggiori rispetto a una strategia di campionamento più semplice come la 2C. All'interno del gruppo 4C, si nota come lo schema nr 38 appartenente al sottogruppo 1r3o (1 campione pre- e 3 campioni post-riduzione) sia quasi sempre selezionato come lo schema migliore del gruppo (Figura2.5a,N_0 150 I_C 2 - 10, N_0 750 I_C 10 - 100, N_0 1500 I_C 10). Questo schema ottiene valori di PC tra i più alti fra tutti i gruppi di schemi (Figura2.5c) e stime di T_C molto vicine al valore vero (T_C medio I_C2=49.23, T_C medio I_C10=53.06, T_C medio I_C100=58.86, Figura2.5b, Tabella2.2-10). Come già verificato nel gruppo C2, spostare il tempo del campionamento pre-riduzione in tempi più antichi (nello schema 38 è posizionato a 50 generazioni nel passato mentre nello schema 39 a 100 generazioni) provoca un abbassamento delle performance in tutti i casi analizzati dovuta alla stima distorta del T_C verso il tempo del campionamento temporale. Tra gli schemi rimanenti, il nr 34, caratterizzato da 4 campioni post-riduzione demografica, ottiene una P_{SS} sempre inferiore a schemi con almeno un campione pre-riduzione (Figura2.5a), suggerendo l'importanza di questo tipo di informazione per ricostruire la dinamica demografica corretta. Gli altri schemi del gruppo C4 (nr 35,40,37,36), nessuno sembra essere supportato in maniera chiara dai valori di P_{SS}, mostrando una forte oscillazione della preferenza tra gli scenari simulativi.

Il passaggio a un campionamento ancora più distribuito nel tempo (8C e 16C) non sembra essere richiesto per ottimizzare la ricostruzione demografica. Infatti, gli schemi 8C e 16C ottengono, in alcuni casi, valori di P_{SS} simile ma non maggiore ad altri schemi di campionamento più semplici contenenti DNA antico (vedi ad esempio $N_0$150 I_C2 o $N_0$150 I_C100 in Figura2.5a) o, negli altri casi, molto inferiori (Figura2.5a).

2.4 DISCUSSIONE

Ricostruire la storia demografica di una popolazione è uno degli aspetti principali a cui si è interessati in genetica di conservazione e in molti altri ambiti, che vanno dall'antropologia molecolare all'epidemiologia. Alcuni metodi statistici di recente sviluppo, permettono di rispondere a questa domanda utilizzando l'informazione combinata proveniente da un campione di sequenze di DNA moderne ed antiche. L'uso del DNA antico è vantaggioso perché contiene al suo interno le informazioni sui processi storici avvenuti nel passato, ma a causa della sua natura non è sempre possibile farne uso. Le numerose difficoltà nell'estrazione e nella genotipizzazione del DNA da campioni biologici degradati, uniti alla generalmente bassa disponibilità dei reperti a disposizione, fanno del DNA antico uno strumento molto costoso da utilizzare. Alla luce di questo, sembra perciò importante valutare quanto, e in che condizioni, il suo uso possa contribuire alle ricostruzioni demografiche. Questo studio di simulazione, sebbene limitato a uno specifico scenario di riduzione demografica, un singolo marcatore genetico e a una specifica metodologia di analisi, conferma con estrema chiarezza il potere del DNA antico nel ricostruire eventi demografici passati. In accordo con quanto verificato da Ramakrishnan et al. 2005, gli schemi che incorporano al loro interno DNA antico hanno migliori prestazioni se confrontati con lo schema avente solo DNA moderno. L'inclusione del dato antico migliora sia la capacità di identificare la riduzione demografica, sia la stima dei suoi parametri più importanti come l'intensità della riduzione e il tempo in cui essa è avvenuta. Quando l'intensità della riduzione demografica è molto forte, ma viene conservata variabilità genetica, lo schema con solo DNA moderno sembra essere sufficiente per identificare l'evento demografico però tende produrre una stima più antica del tempo in cui esso è avvenuto. In questo caso, l'uso del DNA antico permette di raggiungere le stesse performance nell'identificazione della riduzione e di correggere la distorsione nella stima del tempo dell'evento. Inoltre, potendo misurare direttamente il livello di variabilità genetica nel passato, è possibile ricostruire l'evento demografico anche nei casi in cui il livello di polimorfismo nel campione moderno sia estremamente basso. Quando la riduzione demografica non è molto intensa (la popolazione si riduce della metà), includere DNA antico nell'analisi non sembra essere sufficiente per rifiutare l'ipotesi di popolazione costante. Tuttavia, le ricostruzioni prodotte tendono a essere maggiormente non-costanti rispetto allo schema con solo DNA moderno e perciò essere utilizzate come una prima evidenza molto lieve di riduzione demografica.

Contro le attese, gli schemi con due tempi di campionamento hanno performance simili, e in molti casi anche maggiori, rispetto a schemi con un campionamento temporale più distribuito nel tempo. Dividere il campione a disposizione in due parti, una precedente e una successiva all'evento

demografico, sembra essere in assoluto la miglior strategia per ricostruire la dinamica demografica corretta. Dato che lo scenario demografico può essere visto come una transizione istantanea da una popolazione costante pre-riduzione ad una seconda popolazione costante post-riduzione, sembra plausibile che dividere il campione totale in due parti, una prima e una dopo la riduzione, porti ad una stima più precisa delle due dimensioni effettive (pre e post) rispetto a strategie più temporalmente distribuite. Questo schema ottiene ottime performance a patto che il momento del campionamento temporale pre-riduzione non sia troppo lontano dal tempo reale della riduzione demografica. Il tempo della riduzione può essere, infatti, fortemente distorto se il campione pre-riduzione è molto più antico rispetto al tempo reale del declino. Un'ulteriore evidenza di questa distorsione si ha quando si utilizzano schemi con 4 campionamenti temporali. In questo caso, il migliore schema prevede un campione pre-riduzione demografica e tre campioni successivi al declino. Spostando il tempo del campione antico verso tempi maggiori, anche in questo caso, si ottiene un peggioramento delle performance dovute a una stima non corretta del tempo della riduzione. Questo fenomeno è risultato evidente in tulle condizioni testate suggerendo che l'EBSP può in alcuni contesti spostare il tempo della riduzione demografica verso il momento in cui viene campionato il dato antico. Nel caso in cui ci sia la disponibilità di reperti museali lungo tutto un arco temporale, e si abbiano alcune informazioni sul periodo in cui qualche fattore potrebbe aver influenzato la demografia di una specie, la strategia migliore sembra essere quella di procedere alla tipizzazione genetica di un gruppo di individui moderni ed un gruppo di individui antichi che pre-datino la riduzione demografica possibilmente non di troppo tempo in modo da evitare distorsioni nella ricostruzione demografica dovute al campionamento troppo antico. Inoltre il campione moderno non dovrebbe essere preso ad un tempo troppo vicino all'evento demografico ipotetico, in modo che alcuni eventi di coalescenza possano avvenire nell'intervallo di tempo tra il momento del campionamento delle sequenze di DNA e il tempo della riduzione demografica. Più l'intervallo di tempo tende ad essere piccolo, più il numero di eventi di coalescenza al suo interno tende a zero, rendendo di fatto molto difficile (o addirittura impossibile nel caso non ci siano eventi di coalescenza nell'intervallo) identificare un evento demografico con una metodologia basata sulla stima della dimensione effettiva negli intervalli di coalescenza.

Il tipo di campionamento distribuito nel tempo (tipo 4C, 8C e 16C) non sembra aumentare la qualità della ricostruzione demografica rispetto alla strategia a due campioni temporali. Quest'ultima, nel modello di riduzione istantanea studiato, sembra essere più che sufficiente per ottenere una stima accurata delle dimensioni effettive prima e dopo l'evento, dell'intensità della riduzione e del momento temporale in cui è avvenuta. Questo risultato sembra facilitare l'analisi di tutte quelle situazioni in cui sia difficile ottenere campioni antichi da molteplici punti temporali a

causa ad esempio della scarsa disponibilità di reperti museali o del livello di degradazione. Questo risultato indica inoltre che, se non esiste nessuna informazione a priori sulla popolazione, o i campioni pre-riduzione sono troppo antichi, utilizzare schemi di campionamento semplici come il 2C può non essere sufficiente per raggiungere la ricostruzione demografica corretta.

Queste conclusioni, essendo derivate dall'analisi di un modello di riduzione demografica istantanea, non sono estendibili ad altri modelli di riduzione (ad esempio una riduzione lineare o riduzioni multiple) o espansione demografica, per i quali, le performance dei vari schemi di campionamento andrebbero testate specificatamente. Inoltre, i risultati ottenuti non sono esattamente estendibili per riduzioni demografiche avvenute in tempi molto più antichi e anche in questo caso, i risultati dovrebbero essere confermati da ulteriori analisi.

Nel complesso, lo studio ha richiesto l'analisi 37800 dataset di DNA, per un totale di circa 80000 ore/processore utilizzando un cluster linux ad alte prestazioni basato su CPU Intel Xeon a 2.4Ghz. Anche se limitato nelle condizioni esplorate, questo studio risulta difficilmente replicabile in altre condizioni demografiche a causa del notevole sforzo di calcolo necessario. Inoltre, i 42 schemi di campionamento analizzati sono solo un sottoinsieme dei 65535 possibili schemi definibili con 16 posizioni temporali, che se analizzati nella loro totalità avrebbero reso intrattabile il problema a causa dell'enorme mole di calcoli da eseguire. Non è da escludere quindi che esistano schemi di campionamento che non sono stati presi in considerazione che abbiano performance migliori di quelli analizzati. Un approccio che risolverebbe contemporaneamente il problema computazionale e dell'esplorazione dello spazio degli schemi di campionamento possibili, è quello fornito dagli "Algoritmi Genetici" (vedi introduzione in Mitchell 1998), appartenente alla classe degli algoritmi di ottimizzazione ispirato al principio di selezione naturale che regola l'evoluzione biologica. Questo metodo sarebbe facilmente integrabile nel framework utilizzato in questo studio per valutare le performance di uno schema di campionamento, introducendo però una fase di ricerca nello spazio dei possibili schemi che tende a convergere verso l'insieme degli schemi migliori in assoluto per il modello demografico in esame. In questo modo inoltre sarebbero analizzati soltanto un piccolo gruppo di schemi, quelli a performance maggiori e verrebbe perciò evitata l'analisi di una serie di schemi a basse performance di scarso interesse. Quest'approccio perciò sembra essere promettente per estendere i risultati ottenuti in questo studio anche in altri scenari demografici o con un numero maggiore di individui/loci, senza però disporre di grandi capacità elaborative.

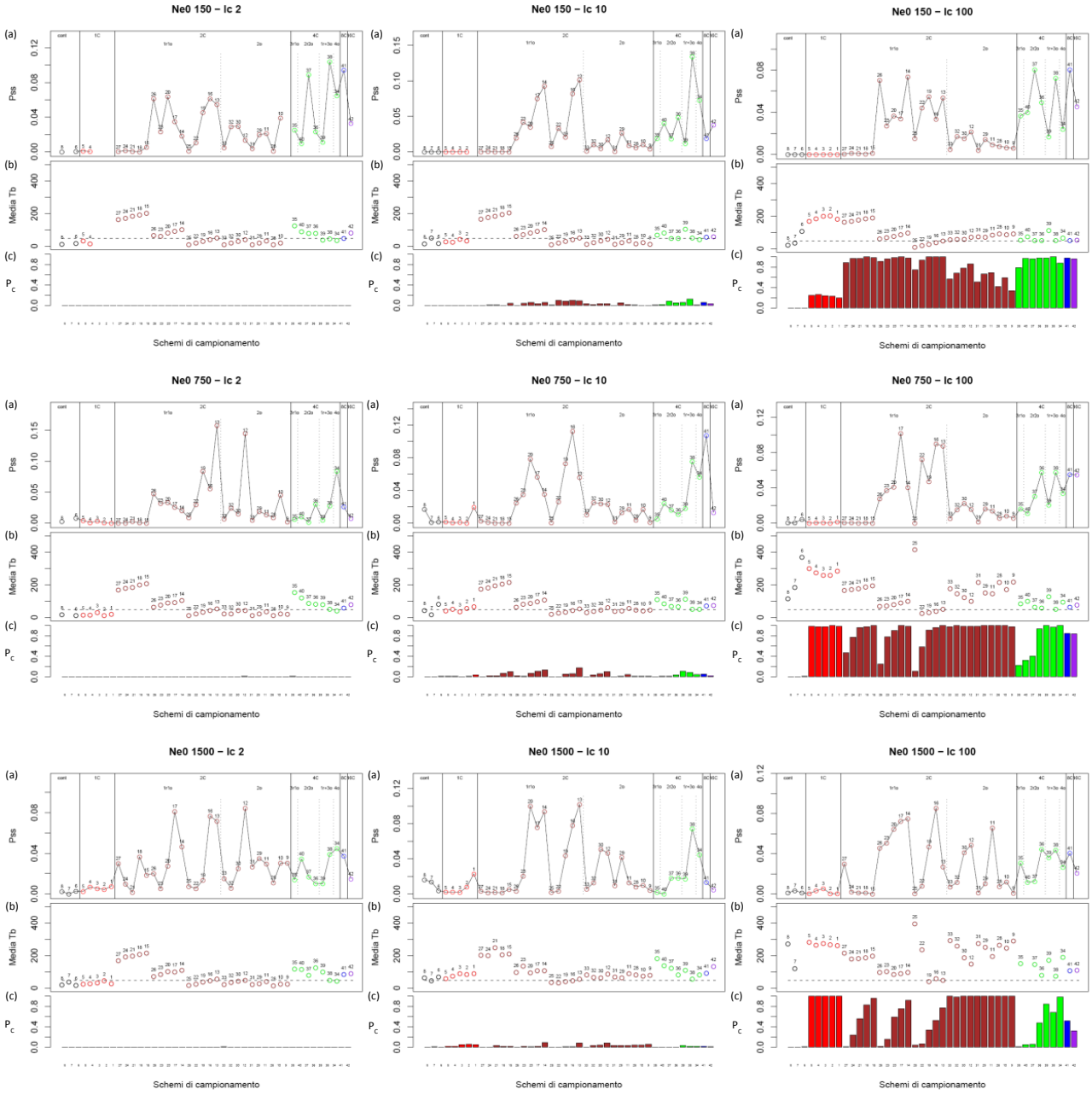

Figura2.11: rappresentazione grafica delle performance nel ricostruire la funzione demografica simulata secondo ogni strategia di campionamento. I grafici sono stati disposti nel seguente modo: N_0 è stata ordinata in modo crescente nelle righe mentre I_C è stato ordinato in modo crescente nelle colonne. Ogni grafico è composto da tre sezioni: (in alto) grafico a dispersione della P_{SS} per ogni schema di campionamento; (intermedia) media del T_C stimato nelle 100 repliche per ogni schema di campionamento; (in basso) istogramma della P_C caratteristica di ogni schema di campionamento. Differenti colori rappresentano i gruppi di schemi di campionamento: gli schemi di controllo in nero, gli schemi con un campione temporale in rosso, due campionamenti temporali in marrone, quattro campionamenti temporali in verde, otto campionamenti temporali in blu e sedici campionamenti in viola. Le aree delimitate sezione più alta in ogni grafico rappresentano la localizzazione dei campionamenti: r (precedente alla riduzione) e o (successivo alla riduzione).

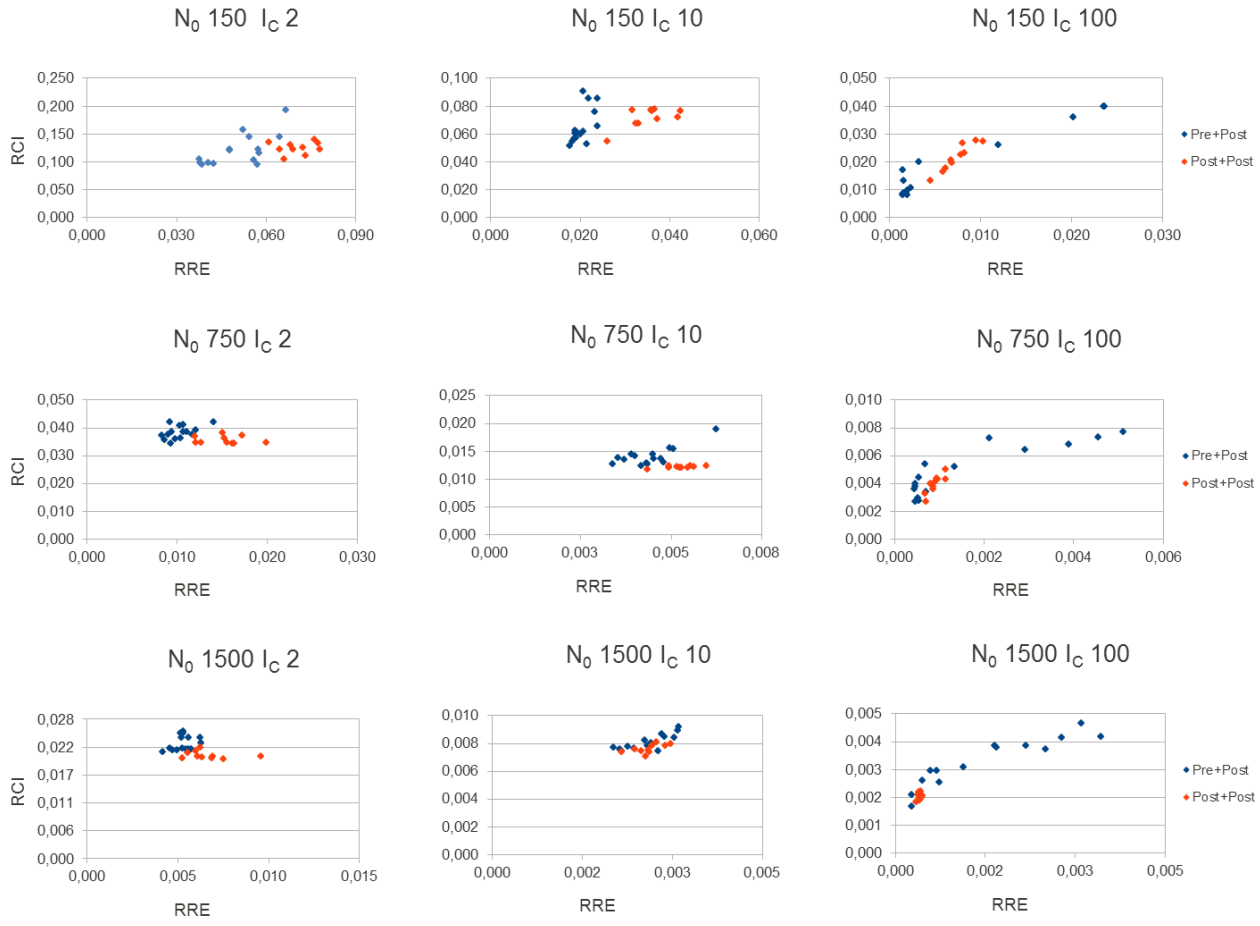

Figura2.12: grafico a dispersione del RRE e RCI degli schemi di campionamento appartenenti al gruppo costituito da 2 campioni temporali. Gli schemi con un campione precedente e uno successivo al tempo della riduzione demografica sono indicati in blu; gli schemi con due campionamenti successivi alla riduzione in rosso.

Tabella2.2: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 150 I_C 2. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,068	0,046	0,098	0,095	0,56	0,50	0,00	-	0,89	0,10	0,93	0,10	0,01	-	14,0	-	167,4	230,5	295,8	219,3	148,4	211,5	0,000	36
7	1	controllo	0,063	0,047	0,106	0,100	0,58	0,50	0,00	-	0,88	0,10	0,92	0,11	0,00	-	-	-	173,8	219,0	271,6	188,1	147,3	188,2	-	38
6	1	controllo	0,061	0,045	0,119	0,102	0,61	0,49	0,00	-	0,87	0,12	0,90	0,13	0,01	-	19,0	-	213,1	267,2	313,0	256,9	182,3	245,7	0,000	33
5	1	1o	0,070	0,046	0,086	0,086	0,51	0,50	0,00	-	0,92	0,07	0,95	0,08	0,01	-	33,0	-	155,1	222,0	308,1	226,7	143,0	216,9	0,001	30
4	1	1o	0,069	0,049	0,099	0,080	0,60	0,49	0,00	-	0,93	0,06	0,95	0,07	0,01	-	16,0	-	170,9	226,2	312,7	206,7	156,7	209,4	0,000	34
3	1	1o	0,082	0,046	0,076	0,086	0,44	0,50	0,00	-	0,90	0,07	0,96	0,07	0,00	-	-	-	99,5	175,3	256,0	184,5	89,0	161,0	-	39
2	1	1o	0,076	0,049	0,093	0,090	0,54	0,50	0,00	-	0,91	0,09	0,95	0,08	0,00	-	-	-	107,5	149,1	231,2	125,8	96,7	137,6	-	40
1	1	1o	0,078	0,049	0,095	0,088	0,55	0,49	0,00	-	0,92	0,06	0,96	0,07	0,00	-	-	-	126,2	199,1	273,5	188,6	113,2	177,7	-	41
27	2	1r1o	0,042	0,033	0,096	0,055	0,73	0,39	0,00	-	0,96	0,04	0,84	0,17	0,04	-	164,0	2,3	172,0	170,1	187,6	162,3	146,3	149,9	0,000	35
24	2	1r1o	0,040	0,030	0,098	0,056	0,75	0,39	0,00	-	0,96	0,05	0,86	0,17	0,05	-	173,0	1,4	169,6	190,9	177,3	184,3	142,8	157,3	0,001	28
21	2	1r1o	0,038	0,028	0,105	0,072	0,79	0,36	0,00	-	0,96	0,05	0,86	0,15	0,01	-	185,0	-	182,1	186,8	187,0	179,9	157,2	162,9	0,001	32
18	2	1r1o	0,039	0,030	0,095	0,053	0,76	0,38	0,00	-	0,96	0,03	0,88	0,22	0,04	-	192,3	1,0	175,8	179,8	184,9	172,3	150,8	154,7	0,000	37
15	2	1r1o	0,038	0,032	0,098	0,054	0,78	0,37	0,00	-	0,96	0,04	0,88	0,17	0,08	-	203,8	2,5	198,8	193,6	217,5	186,4	177,2	185,2	0,005	25
26	2	1r1o	0,052	0,037	0,158	0,144	0,76	0,38	0,00	-	0,89	0,27	0,85	0,24	0,06	-	66,2	11,2	180,2	211,3	227,7	182,5	150,2	175,2	0,061	7
23	2	1r1o	0,056	0,040	0,103	0,084	0,64	0,45	0,00	-	0,89	0,16	0,87	0,16	0,03	-	62,3	22,0	179,1	232,6	261,9	225,3	152,6	208,5	0,023	16
20	2	1r1o	0,048	0,036	0,121	0,089	0,74	0,38	0,00	-	0,93	0,34	0,89	0,31	0,06	-	81,3	1,5	147,8	146,5	192,1	135,5	130,2	136,5	0,064	5
17	2	1r1o	0,048	0,031	0,122	0,065	0,75	0,36	0,00	-	0,90	0,17	0,85	0,17	0,05	-	92,4	0,9	164,1	187,8	196,6	183,9	140,9	168,3	0,035	11
14	2	1r1o	0,057	0,034	0,095	0,073	0,63	0,42	0,00	-	0,89	0,17	0,87	0,16	0,02	-	102,0	1,4	107,0	148,7	148,0	154,8	95,4	136,5	0,018	20
25	2	1r1o	0,067	0,048	0,193	0,391	0,73	0,43	0,00	-	0,90	0,19	0,91	0,17	0,06	-	11,0	0,9	176,6	240,4	271,1	248,8	153,8	225,6	0,001	29
22	2	1r1o	0,064	0,047	0,145	0,137	0,78	0,40	0,00	-	0,89	0,20	0,86	0,19	0,03	-	21,7	1,2	139,3	166,1	214,7	131,5	116,8	137,9	0,010	23
19	2	1r1o	0,054	0,044	0,145	0,091	0,80	0,38	0,00	-	1,23	2,52	1,06	1,84	0,08	-	30,6	1,5	166,8	172,6	227,6	150,9	147,7	152,8	0,046	9
16	2	1r1o	0,057	0,042	0,123	0,098	0,76	0,41	0,00	-	0,96	0,21	0,90	0,19	0,14	-	41,6	1,5	135,8	158,6	215,7	160,0	126,7	158,6	0,062	6
13	2	1r1o	0,058	0,040	0,116	0,089	0,67	0,42	0,00	-	0,92	0,33	0,88	0,29	0,06	-	51,3	0,8	147,3	179,0	210,6	154,1	124,0	150,5	0,055	8
33	2	2o	0,072	0,049	0,125	0,091	0,71	0,44	0,00	-	0,90	0,14	0,93	0,11	0,02	-	11,0	0,0	106,2	140,8	195,0	130,8	95,0	129,6	0,005	26
32	2	2o	0,068	0,045	0,130	0,078	0,74	0,42	0,00	-	0,96	0,42	0,90	0,31	0,08	-	21,5	1,4	109,1	141,8	204,4	134,7	100,4	134,6	0,029	14
30	2	2o	0,061	0,044	0,136	0,086	0,78	0,40	0,00	-	0,94	0,25	0,89	0,22	0,07	-	30,6	0,8	146,6	194,4	232,1	160,3	127,5	153,7	0,029	13
12	2	2o	0,066	0,040	0,105	0,084	0,68	0,45	0,00	-	0,96	0,41	0,92	0,30	0,03	-	41,7	0,6	100,2	133,7	178,6	132,6	91,1	124,9	0,014	21
31	2	2o	0,078	0,051	0,123	0,078	0,71	0,45	0,00	-	0,92	0,08	0,93	0,08	0,01	-	11,0	-	105,6	164,6	227,3	166,0	92,5	146,0	0,004	27
29	2	2o	0,064	0,044	0,123	0,078	0,78	0,40	0,00	-	0,98	0,55	0,92	0,43	0,05	-	21,8	1,9	136,8	182,9	233,2	211,1	129,8	190,0	0,020	19
11	2	2o	0,073	0,043	0,110	0,074	0,69	0,44	0,00	-	0,91	0,17	0,87	0,16	0,05	-	34,4	5,0	106,1	162,6	209,0	179,0	95,3	153,8	0,021	18
28	2	2o	0,077	0,045	0,134	0,104	0,69	0,45	0,00	-	0,90	0,15	0,92	0,11	0,01	-	10,0	-	91,8	137,0	210,8	117,6	79,8	119,5	0,001	31
10	2	2o	0,069	0,044	0,122	0,121	0,71	0,45	0,00	-	1,21	1,33	1,04	0,99	0,06	-	21,8	1,6	96,9	127,2	209,1	197,6	104,6	175,8	0,039	10
9	2	2o	0,076	0,047	0,140	0,122	0,72	0,44	0,00	-	0,89	0,11	0,92	0,10	0,00	-	-	-	93,0	134,1	190,9	123,5	80,5	118,2	-	42
35	4	3r1o	0,033	0,025	0,107	0,077	0,80	0,37	0,00	-	0,93	0,13	0,87	0,20	0,12	-	124,8	53,1	210,1	202,6	211,2	193,8	188,8	189,3	0,025	15
40	4	2r2o	0,036	0,027	0,097	0,063	0,77	0,36	0,00	-	0,94	0,18	0,92	0,44	0,09	-	90,9	76,1	174,2	173,8	178,1	168,1	162,1	172,2	0,009	24
37	4	2r2o	0,034	0,026	0,114	0,156	0,80	0,36	0,00	-	0,91	0,35	0,87	0,43	0,11	-	80,2	64,0	176,1	154,8	175,2	152,9	158,9	151,1	0,089	3
36	4	2r2o	0,037	0,030	0,097	0,057	0,77	0,39	0,00	-	0,95	0,22	0,91	0,34	0,09	-	81,0	74,1	184,4	175,9	197,8	171,6	169,2	161,7	0,023	17
39	4	1r3o	0,039	0,032	0,086	0,052	0,76	0,39	0,00	-	0,93	0,18	0,90	0,21	0,06	-	38,5	8,6	162,9	161,7	176,7	154,9	148,1	148,7	0,011	22
38	4	1r3o	0,055	0,038	0,132	0,103	0,75	0,38	0,00	-	1,05	0,34	0,95	0,31	0,11	-	46,4	8,8	131,8	178,3	198,6	167,4	122,4	151,7	0,104	1
34	4	4o	0,059	0,041	0,107	0,068	0,76	0,41	0,00	-	0,98	0,35	0,91	0,30	0,07	-	36,6	5,3	107,5	122,6	160,5	103,9	96,5	108,3	0,065	4
41	8	3r5o	0,033	0,027	0,103	0,091	0,78	0,36	0,00	-	1,03	0,73	1,00	0,80	0,16	-	48,9	16,3	193,4	169,7	211,4	174,3	190,7	172,7	0,095	2
42	16	6r10o	0,028	0,022	0,114	0,063	0,91	0,25	0,00	-	0,97	0,33	0,95	0,42	0,13	-	81,6	82,0	228,5	191,1	232,1	191,2	217,2	203,1	0,033	12

Tabella2.3: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 150 I_C 10. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,027	0,010	0,102	0,077	0,77	0,19	0,00	-	0,94	0,09	0,88	0,13	0,04	-	15,8	3,4	1346,7	709,7	1273,4	693,7	1183,5	637,3	0,000	38
7	1	controllo	0,026	0,010	0,099	0,057	0,75	0,20	0,00	-	0,94	0,08	0,89	0,14	0,04	-	50,5	21,9	1423,8	709,4	1348,3	706,1	1256,1	640,3	0,000	39
6	1	controllo	0,030	0,019	0,114	0,086	0,73	0,21	0,00	-	0,92	0,11	0,86	0,17	0,02	-	19,5	13,4	1442,2	782,6	1352,4	711,6	1207,2	622,5	0,000	41
5	1	1o	0,031	0,027	0,077	0,047	0,84	0,23	0,00	-	0,97	0,07	0,97	0,09	0,04	-	28,8	11,2	634,3	439,4	687,1	441,7	638,0	484,2	0,000	37
4	1	1o	0,034	0,034	0,070	0,034	0,80	0,28	0,00	-	0,97	0,06	0,97	0,09	0,04	-	26,8	16,0	610,8	433,8	671,5	435,3	607,9	482,0	0,000	32
3	1	1o	0,040	0,035	0,083	0,082	0,79	0,31	0,00	-	0,96	0,08	0,96	0,09	0,02	-	43,5	17,7	494,6	401,2	585,0	383,8	485,3	420,1	0,000	36
2	1	1o	0,043	0,038	0,075	0,072	0,75	0,35	0,00	-	0,96	0,09	0,97	0,11	0,03	-	35,0	22,1	493,1	411,2	614,0	404,3	502,3	479,2	0,000	33
1	1	1o	0,038	0,035	0,067	0,044	0,79	0,33	0,00	-	0,96	0,06	0,97	0,07	0,00	-	-	-	475,5	354,6	558,0	326,5	472,4	372,4	-	42
27	2	1r1o	0,019	0,009	0,056	0,023	0,82	0,18	0,00	-	1,01	0,04	1,24	0,68	0,33	-	167,5	6,2	887,9	409,4	901,3	423,1	1046,8	525,2	0,000	35
24	2	1r1o	0,019	0,009	0,061	0,033	0,78	0,23	0,01	-	1,01	0,05	1,51	1,32	0,34	-	177,7	8,3	788,9	425,5	795,9	429,6	1020,5	611,3	0,001	29
21	2	1r1o	0,019	0,007	0,057	0,036	0,75	0,24	0,01	-	1,00	0,04	1,43	1,25	0,30	-	184,9	4,1	791,4	434,2	797,2	444,0	999,9	614,5	0,000	31
18	2	1r1o	0,021	0,017	0,052	0,018	0,73	0,27	0,00	-	1,01	0,04	1,36	0,90	0,32	-	195,2	4,7	644,8	335,1	666,1	332,1	825,1	468,3	0,000	34
15	2	1r1o	0,018	0,008	0,052	0,018	0,78	0,25	0,04	-	1,01	0,05	1,92	2,22	0,38	-	204,4	5,0	614,2	336,9	621,2	342,1	923,6	588,1	0,000	40
26	2	1r1o	0,024	0,022	0,066	0,046	0,82	0,26	0,00	-	1,91	1,87	1,79	1,73	0,43	-	62,7	3,2	691,5	496,9	1049,8	591,2	956,1	622,8	0,020	15
23	2	1r1o	0,019	0,012	0,061	0,040	0,83	0,21	0,04	-	2,49	5,55	2,45	5,53	0,43	-	72,9	3,4	578,2	356,2	940,2	714,2	922,8	721,1	0,042	8
20	2	1r1o	0,019	0,011	0,062	0,036	0,80	0,25	0,06	-	2,61	3,51	2,61	3,49	0,45	-	83,3	5,0	537,6	362,5	936,8	629,3	939,0	647,7	0,035	11
17	2	1r1o	0,018	0,010	0,054	0,024	0,80	0,24	0,03	-	2,25	3,17	2,26	3,10	0,51	-	93,1	3,9	579,5	400,5	901,9	576,1	907,4	583,8	0,075	5
14	2	1r1o	0,021	0,016	0,061	0,040	0,78	0,27	0,06	-	3,48	7,91	3,46	7,23	0,45	-	102,0	6,1	473,4	323,3	861,3	576,9	857,4	632,1	0,093	3
25	2	1r1o	0,024	0,011	0,085	0,056	0,87	0,13	0,01	-	1,07	0,22	1,05	0,24	0,24	-	11,4	1,5	909,6	567,9	983,7	642,5	965,4	652,5	0,007	24
22	2	1r1o	0,022	0,009	0,086	0,066	0,92	0,10	0,10	-	5,66	12,53	5,63	12,79	0,47	-	21,0	1,1	514,0	424,0	952,0	627,6	923,9	642,3	0,033	12
19	2	1r1o	0,023	0,014	0,076	0,057	0,90	0,12	0,08	-	4,39	8,17	4,21	7,91	0,48	-	31,3	1,4	425,3	359,9	857,3	650,3	807,9	595,6	0,021	14
16	2	1r1o	0,021	0,016	0,091	0,224	0,89	0,17	0,10	-	3,91	5,81	3,78	5,64	0,56	-	41,9	2,4	487,9	407,9	1031,2	682,6	991,9	696,1	0,082	4
13	2	1r1o	0,020	0,015	0,060	0,038	0,83	0,24	0,09	-	5,00	13,54	4,88	13,25	0,56	-	51,9	2,7	409,1	335,7	917,9	739,0	893,8	745,7	0,101	2
33	2	2o	0,036	0,032	0,077	0,051	0,83	0,27	0,03	-	1,14	0,36	1,12	0,37	0,19	-	13,8	6,6	494,1	405,7	644,9	455,7	567,5	519,8	0,001	28
32	2	2o	0,032	0,024	0,077	0,044	0,86	0,18	0,02	-	1,75	2,61	1,70	2,55	0,23	-	24,1	7,5	401,7	324,8	537,3	389,5	510,7	419,2	0,010	21
30	2	2o	0,032	0,030	0,067	0,043	0,84	0,23	0,03	-	1,87	2,31	1,80	2,43	0,31	-	32,8	4,5	382,1	328,1	595,2	437,8	544,5	480,4	0,005	26
12	2	2o	0,026	0,023	0,055	0,030	0,81	0,21	0,03	-	2,13	3,08	2,13	3,22	0,34	-	43,4	5,0	379,9	312,9	591,7	401,0	580,6	453,0	0,017	19
31	2	2o	0,036	0,029	0,076	0,046	0,80	0,27	0,00	-	1,07	0,24	1,06	0,25	0,17	-	12,6	3,8	436,0	357,2	535,1	397,8	478,6	420,2	0,001	30
29	2	2o	0,037	0,034	0,071	0,045	0,82	0,26	0,05	-	2,51	3,22	2,38	3,29	0,21	-	23,5	4,2	307,4	291,7	567,4	462,5	519,9	570,1	0,027	13
11	2	2o	0,033	0,026	0,067	0,042	0,81	0,23	0,02	-	2,14	3,66	2,05	3,62	0,22	-	32,6	2,6	315,0	258,9	527,1	436,1	487,4	460,0	0,009	23
28	2	2o	0,037	0,035	0,078	0,047	0,85	0,22	0,01	-	1,24	1,31	1,26	1,44	0,17	-	16,8	11,0	428,7	401,4	558,4	530,6	533,1	604,6	0,006	25
10	2	2o	0,042	0,036	0,076	0,056	0,81	0,30	0,00	-	1,64	1,63	1,51	1,57	0,18	-	24,0	6,1	314,3	288,5	515,4	364,4	422,9	412,6	0,010	22
9	2	2o	0,042	0,038	0,072	0,040	0,82	0,27	0,00	-	1,17	0,60	1,15	0,59	0,14	-	12,6	2,6	350,9	294,7	492,6	338,4	411,3	374,9	0,004	27
35	4	3r1o	0,022	0,015	0,069	0,051	0,83	0,30	0,01	-	2,30	4,33	2,19	3,95	0,35	-	68,5	46,2	779,7	578,2	997,2	559,0	928,0	522,8	0,019	17
40	4	2r2o	0,018	0,009	0,059	0,025	0,81	0,26	0,02	-	1,65	2,00	1,88	2,50	0,46	-	82,7	55,6	725,3	399,8	957,7	542,6	1019,4	598,5	0,040	9
37	4	2r2o	0,017	0,010	0,056	0,028	0,85	0,26	0,08	-	5,27	11,31	5,18	10,64	0,58	-	50,2	4,4	615,3	490,5	1006,3	564,7	1016,5	570,3	0,018	18
36	4	2r2o	0,017	0,011	0,057	0,029	0,85	0,24	0,05	-	5,12	13,45	5,21	13,78	0,56	-	49,3	8,2	530,8	412,3	906,0	493,0	921,6	496,0	0,048	7
39	4	1r3o	0,017	0,012	0,053	0,032	0,77	0,24	0,06	-	2,83	5,67	3,75	6,50	0,51	-	105,2	79,6	460,3	365,1	664,2	444,9	865,9	564,2	0,012	20
38	4	1r3o	0,020	0,019	0,064	0,053	0,87	0,21	0,12	-	8,35	15,78	8,05	15,56	0,61	-	50,5	2,6	312,0	317,5	932,6	669,5	890,0	690,9	0,134	1
34	4	4o	0,030	0,020	0,072	0,061	0,82	0,17	0,01	-	2,31	2,88	2,27	2,89	0,31	-	40,6	6,6	300,2	282,0	472,5	396,9	459,6	424,9	0,073	6
41	8	3r5o	0,019	0,011	0,060	0,033	0,79	0,27	0,06	-	3,27	5,13	3,37	5,19	0,50	-	56,4	28,9	495,5	384,9	821,5	502,5	854,1	575,3	0,019	16
42	16	6r10o	0,018	0,014	0,057	0,035	0,85	0,24	0,03	-	3,02	4,48	3,24	4,94	0,59	-	59,2	27,3	615,3	476,1	1007,9	587,8	1061,3	679,3	0,038	10

Tabella2.4: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 150 I_C 100. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,028	0,015	0,057	0,040	0,86	0,11	0,00	-	1,00	0,03	0,97	0,12	0,09	-	23,8	13,4	15290,0	3428,5	15340,7	3380,3	14659,1	3219,4	0,000	36
7	1	controllo	0,028	0,017	0,054	0,048	0,86	0,07	0,00	-	1,01	0,02	0,99	0,13	0,07	-	36,0	21,0	14927,1	4184,6	15010,3	4183,4	14429,9	2750,9	0,000	42
6	1	controllo	0,031	0,021	0,063	0,071	0,86	0,12	0,00	-	1,01	0,02	0,97	0,18	0,08	-	108,9	130,7	15455,7	5033,8	15536,8	4989,7	14618,7	3574,4	0,000	37
5	1	1o	0,007	0,006	0,021	0,009	0,50	0,27	0,25	-	1,57	0,44	5,86	5,03	0,83	-	168,8	147,4	1941,7	624,8	2973,3	1075,9	9307,5	6061,6	0,000	40
4	1	1o	0,008	0,006	0,021	0,008	0,54	0,26	0,26	-	1,46	0,35	6,07	5,51	0,72	-	184,8	151,6	1697,8	590,8	2405,9	786,3	8154,3	5433,7	0,000	41
3	1	1o	0,007	0,005	0,020	0,008	0,62	0,18	0,24	-	1,42	0,37	6,52	5,42	0,74	-	201,3	132,7	1428,3	423,2	2003,4	680,1	8083,3	5525,4	0,000	38
2	1	1o	0,009	0,008	0,023	0,017	0,62	0,21	0,23	-	1,34	0,29	6,67	7,09	0,65	-	204,0	134,7	1307,5	513,0	1737,4	679,2	7010,5	5855,1	0,000	39
1	1	1o	0,011	0,013	0,022	0,014	0,61	0,19	0,20	-	1,29	0,27	5,53	6,87	0,54	-	183,3	141,1	1281,7	485,0	1669,9	656,1	5794,0	5814,2	0,000	35
27	2	1r1o	0,002	0,001	0,011	0,006	0,73	0,17	0,89	-	1,20	0,29	17,02	8,16	1,00	-	164,1	13,3	981,6	409,8	1164,5	498,2	14388,5	3687,3	0,000	34
24	2	1r1o	0,002	0,001	0,009	0,005	0,81	0,15	0,97	-	1,10	0,08	26,86	16,17	1,00	-	169,1	9,9	637,4	232,3	707,7	292,2	14477,1	3737,1	0,001	31
21	2	1r1o	0,002	0,001	0,010	0,008	0,80	0,15	0,97	-	1,06	0,07	27,23	14,29	1,00	-	176,9	13,7	613,2	270,2	657,6	319,2	13913,2	3751,5	0,001	32
18	2	1r1o	0,002	0,001	0,008	0,003	0,82	0,15	1,00	-	1,03	0,06	32,65	18,46	1,00	-	184,0	10,7	474,9	168,1	496,1	192,5	13240,9	3794,2	0,001	33
15	2	1r1o	0,002	0,001	0,009	0,004	0,86	0,11	0,99	-	1,01	0,05	36,75	20,25	1,00	-	190,6	11,9	443,1	171,2	453,2	188,1	13638,3	3468,7	0,001	30
26	2	1r1o	0,001	0,001	0,017	0,011	0,94	0,13	0,91	-	41,04	22,80	40,57	22,73	1,00	-	62,3	12,1	453,2	335,9	14078,5	3382,6	14033,2	3348,5	0,071	5
23	2	1r1o	0,002	0,001	0,013	0,010	0,94	0,14	0,96	-	53,25	28,72	52,24	29,99	1,00	-	69,2	4,6	352,4	350,8	14001,6	3684,6	13676,4	3723,9	0,027	16
20	2	1r1o	0,002	0,001	0,009	0,005	0,92	0,15	0,99	-	57,71	29,98	55,95	26,34	1,00	-	78,2	5,5	273,4	111,9	13268,1	3399,0	13101,8	3412,3	0,037	13
17	2	1r1o	0,001	0,001	0,008	0,003	0,95	0,09	1,00	-	72,28	42,44	72,92	44,42	1,00	-	88,2	5,8	236,9	102,8	13823,7	3259,6	13985,3	3613,8	0,034	14
14	2	1r1o	0,001	0,001	0,009	0,005	0,94	0,10	0,98	-	57,35	29,15	62,27	33,26	1,00	-	97,1	5,9	249,8	96,0	12207,1	2625,7	13178,9	3259,0	0,074	3
25	2	1r1o	0,024	0,012	0,040	0,025	0,72	0,19	0,75	-	1,42	1,48	1,43	1,58	0,67	-	11,2	2,1	10795,1	3224,9	13505,0	3113,4	13444,5	3312,0	0,015	21
22	2	1r1o	0,024	0,012	0,040	0,024	0,82	0,21	0,94	-	99,81	95,30	94,39	88,07	1,00	-	19,9	1,9	620,6	828,2	13378,1	3595,7	12864,7	3716,6	0,044	10
19	2	1r1o	0,020	0,010	0,036	0,020	0,87	0,18	1,00	-	166,25	94,35	159,27	90,05	1,00	-	29,7	2,6	105,9	44,6	14494,1	3605,3	13962,5	3609,7	0,055	6
16	2	1r1o	0,012	0,007	0,026	0,014	0,93	0,15	1,00	-	143,25	82,57	140,80	83,58	1,00	-	39,4	3,4	117,0	47,8	13785,7	3217,5	13539,7	3382,5	0,033	15
13	2	1r1o	0,003	0,009	0,020	0,023	0,94	0,14	1,00	-	151,86	129,51	146,11	127,76	1,00	-	49,2	4,5	123,8	59,9	14299,2	3771,1	13764,5	3546,4	0,054	7
33	2	2o	0,009	0,007	0,028	0,014	0,73	0,27	0,57	-	19,52	48,90	24,77	50,75	0,88	-	57,7	66,1	1434,9	1895,6	6511,2	5009,2	9907,4	5551,4	0,005	28
32	2	2o	0,008	0,005	0,023	0,009	0,72	0,27	0,68	-	50,10	58,59	58,24	61,29	0,96	-	59,8	64,0	435,9	552,8	6148,9	4179,0	8721,0	5169,8	0,017	19
30	2	2o	0,006	0,005	0,018	0,008	0,76	0,22	0,78	-	58,17	70,85	65,12	75,36	0,94	-	59,3	39,2	280,6	286,0	7267,0	5199,5	8714,3	5650,4	0,015	22
12	2	2o	0,004	0,005	0,013	0,008	0,77	0,21	0,86	-	59,97	62,17	69,33	65,84	0,95	-	71,5	35,4	283,4	293,5	7644,4	4868,6	9810,5	5520,7	0,021	18
31	2	2o	0,008	0,007	0,027	0,013	0,74	0,24	0,51	-	20,41	37,22	28,30	41,02	0,88	-	75,2	78,7	920,2	851,7	5448,6	4284,6	9481,1	6443,4	0,004	29
29	2	2o	0,007	0,005	0,021	0,009	0,78	0,20	0,67	-	46,62	63,85	57,25	67,63	0,92	-	71,6	70,7	386,0	477,8	5543,3	3819,3	8774,8	5722,0	0,014	23
11	2	2o	0,006	0,005	0,016	0,009	0,72	0,23	0,69	-	41,02	82,25	51,01	84,45	0,88	-	84,3	55,8	434,2	444,3	4922,4	3868,1	8150,2	5651,8	0,009	24
28	2	2o	0,008	0,009	0,023	0,016	0,73	0,22	0,42	-	18,42	41,39	26,61	44,49	0,83	-	92,8	104,8	825,5	575,7	3966,7	3891,6	8336,8	6247,1	0,008	25
10	2	2o	0,007	0,007	0,020	0,009	0,75	0,21	0,59	-	39,06	60,11	50,29	66,32	0,87	-	87,5	77,0	453,2	451,8	4512,2	4169,4	8085,5	5942,5	0,006	26
9	2	2o	0,010	0,012	0,027	0,027	0,69	0,22	0,34	-	11,70	26,03	17,59	30,07	0,70	-	92,8	92,1	805,3	525,3	3155,1	2816,8	6142,4	4811,1	0,006	27
35	4	3r1o	0,005	0,009	0,028	0,025	0,92	0,20	0,79	-	115,79	134,81	114,29	132,93	0,98	-	53,5	13,9	535,3	1565,7	13172,1	3480,5	13374,1	3416,0	0,037	12
40	4	2r2o	0,003	0,003	0,018	0,011	0,92	0,16	0,97	-	60,29	48,34	61,14	49,38	1,00	-	74,9	25,9	317,2	228,0	12537,1	3444,9	13115,5	3417,1	0,040	11
37	4	2r2o	0,002	0,002	0,013	0,007	0,88	0,23	0,96	-	132,96	71,03	131,98	71,64	1,00	-	50,6	3,0	140,9	190,0	13290,4	3430,8	13279,1	3572,4	0,081	1
36	4	2r2o	0,002	0,004	0,019	0,012	0,90	0,18	0,98	-	134,09	69,70	131,16	67,57	1,00	-	51,1	4,1	124,3	64,6	13396,2	3131,9	13268,0	3212,3	0,049	8
39	4	1r3o	0,002	0,002	0,011	0,009	0,83	0,19	0,98	-	65,11	95,10	86,46	96,13	1,00	-	114,4	62,1	256,7	166,8	6933,4	5588,6	12809,3	4158,0	0,017	20
38	4	1r3o	0,002	0,002	0,012	0,006	0,77	0,27	1,00	-	162,99	87,32	163,12	87,87	1,00	-	50,5	2,6	95,7	53,6	13165,0	3866,5	13235,6	4033,5	0,072	4
34	4	4o	0,005	0,004	0,014	0,005	0,76	0,24	0,88	-	78,36	68,07	87,49	71,68	0,97	-	68,3	60,7	208,3	257,3	7819,9	4876,2	9603,5	5161,8	0,024	17
41	8	3r5o	0,003	0,003	0,016	0,011	0,82	0,26	0,98	-	141,81	77,87	140,64	69,40	1,00	-	51,9	7,4	116,8	101,7	12431,2	3809,2	12690,4	3971,6	0,081	2
42	16	6r10o	0,004	0,005	0,015	0,010	0,85	0,23	0,96	-	130,44	88,40	132,32	89,50	1,00	-	53,9	10,9	138,5	105,7	12615,7	3809,4	12930,0	3792,6	0,046	9

Tabella2.5: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 750 I_C 2. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,014	0,017	0,039	0,018	0,95	0,15	0,00	-	0,93	0,10	0,89	0,16	0,03	-	19,7	9,0	1345,0	817,9	1294,1	753,8	1196,6	739,1	0,003	31
7	1	controllo	0,012	0,010	0,044	0,014	0,98	0,05	0,00	-	0,90	0,09	0,83	0,14	0,00	-	-	-	1252,7	614,0	1126,1	551,0	1024,3	500,7	-	42
6	1	controllo	0,011	0,009	0,040	0,013	0,98	0,06	0,00	-	0,93	0,09	0,89	0,14	0,02	-	13,0	1,4	1336,7	717,3	1257,2	697,2	1190,6	664,2	0,008	23
5	1	1o	0,017	0,024	0,035	0,014	0,92	0,22	0,00	-	0,97	0,09	0,96	0,13	0,05	-	17,0	5,0	1069,9	637,3	1114,5	624,0	1044,0	667,2	0,004	29
4	1	1o	0,017	0,025	0,034	0,011	0,92	0,22	0,00	-	0,97	0,08	0,95	0,11	0,05	-	17,0	5,2	1063,5	634,5	1099,0	627,2	1030,1	667,2	0,001	35
3	1	1o	0,013	0,014	0,036	0,009	0,96	0,12	0,00	-	0,95	0,09	0,93	0,13	0,03	-	32,3	12,0	1022,5	526,3	1000,0	538,7	968,1	554,2	0,003	30
2	1	1o	0,016	0,020	0,034	0,009	0,93	0,18	0,00	-	0,96	0,06	0,93	0,09	0,01	-	12,0	-	938,4	570,1	931,7	541,9	875,7	537,2	0,000	36
1	1	1o	0,018	0,024	0,034	0,008	0,93	0,15	0,00	-	0,95	0,07	0,93	0,10	0,02	-	20,0	9,9	912,3	598,9	907,9	609,2	872,4	611,8	0,000	38
27	2	1r1o	0,009	0,009	0,034	0,012	0,94	0,18	0,00	-	0,98	0,03	0,93	0,20	0,12	-	169,8	5,1	1352,8	653,2	1331,0	650,3	1235,5	619,1	0,000	40
24	2	1r1o	0,010	0,008	0,036	0,013	0,92	0,19	0,00	-	0,99	0,03	0,94	0,20	0,13	-	179,6	12,2	1227,7	616,8	1220,2	613,7	1153,7	611,2	0,000	39
21	2	1r1o	0,009	0,008	0,038	0,021	0,96	0,14	0,00	-	0,98	0,04	0,93	0,28	0,06	-	184,5	1,2	1284,7	620,1	1265,6	620,4	1166,1	537,6	0,002	32
18	2	1r1o	0,008	0,007	0,037	0,015	0,95	0,15	0,00	-	0,99	0,04	0,95	0,29	0,17	-	201,1	12,2	1292,2	566,2	1280,4	554,5	1205,9	529,2	0,000	37
15	2	1r1o	0,009	0,007	0,036	0,012	0,94	0,17	0,00	-	0,99	0,05	0,99	0,35	0,14	-	209,1	7,9	1320,9	649,0	1306,8	638,4	1263,5	624,9	0,000	41
26	2	1r1o	0,011	0,010	0,039	0,018	0,93	0,19	0,00	-	1,06	0,50	1,02	0,53	0,18	-	64,9	6,4	1158,3	585,8	1206,0	652,0	1155,2	644,6	0,047	6
23	2	1r1o	0,011	0,010	0,039	0,013	0,91	0,21	0,00	-	1,03	0,34	0,98	0,40	0,13	-	76,6	8,7	1171,9	639,7	1190,3	660,3	1099,6	571,7	0,032	9
20	2	1r1o	0,010	0,009	0,041	0,021	0,95	0,15	0,00	-	1,04	0,54	0,99	0,65	0,09	-	89,4	6,6	1334,0	690,5	1323,6	668,7	1232,2	664,0	0,033	8
17	2	1r1o	0,011	0,008	0,041	0,032	0,93	0,19	0,00	-	1,04	0,21	0,99	0,34	0,21	-	93,3	2,2	1162,1	625,3	1199,7	644,8	1120,8	646,0	0,026	14
14	2	1r1o	0,010	0,010	0,036	0,010	0,91	0,20	0,00	-	1,10	0,66	1,06	0,77	0,21	-	105,2	4,3	1115,4	584,8	1154,8	581,6	1096,3	576,0	0,020	16
25	2	1r1o	0,014	0,022	0,042	0,015	0,95	0,18	0,00	-	0,93	0,11	0,88	0,16	0,04	-	13,5	3,8	1311,8	690,9	1264,9	635,9	1136,2	629,4	0,009	22
22	2	1r1o	0,009	0,007	0,042	0,015	0,98	0,11	0,00	-	0,97	0,21	0,92	0,25	0,13	-	23,7	4,9	1374,1	641,5	1319,8	636,5	1232,1	617,1	0,030	11
19	2	1r1o	0,009	0,008	0,039	0,014	0,97	0,10	0,00	-	1,28	1,08	1,25	1,12	0,34	-	33,2	5,8	1152,1	615,4	1288,0	653,8	1224,8	633,0	0,084	3
16	2	1r1o	0,012	0,013	0,039	0,031	0,92	0,19	0,00	-	1,06	0,26	1,03	0,30	0,24	-	44,3	6,8	1103,7	652,5	1161,7	663,0	1099,5	641,0	0,055	5
13	2	1r1o	0,012	0,010	0,037	0,012	0,93	0,18	0,00	-	1,15	0,59	1,10	0,60	0,25	-	54,3	4,1	979,5	506,7	1081,5	600,6	1030,9	597,1	0,158	1
33	2	2o	0,013	0,019	0,035	0,007	0,95	0,18	0,00	-	1,00	0,13	0,98	0,16	0,12	-	23,6	20,0	1104,6	571,1	1136,7	594,6	1104,0	635,8	0,007	25
32	2	2o	0,012	0,013	0,035	0,009	0,96	0,12	0,00	-	1,04	0,43	1,02	0,45	0,11	-	23,3	3,4	1068,5	581,2	1093,1	604,0	1050,9	599,2	0,024	15
30	2	2o	0,015	0,012	0,038	0,018	0,92	0,18	0,00	-	1,02	0,43	0,98	0,47	0,14	-	43,1	20,0	927,3	547,0	898,3	536,1	866,5	568,9	0,015	18
12	2	2o	0,012	0,009	0,037	0,016	0,95	0,15	0,01	-	1,26	1,41	1,23	1,42	0,20	-	44,7	6,8	911,7	454,4	965,2	487,3	930,5	490,5	0,145	2
31	2	2o	0,015	0,016	0,036	0,012	0,93	0,18	0,00	-	0,96	0,14	0,93	0,17	0,08	-	14,0	6,0	970,3	544,6	957,1	568,4	915,3	577,8	0,005	28
29	2	2o	0,016	0,018	0,035	0,009	0,92	0,19	0,00	-	0,99	0,17	0,96	0,20	0,08	-	23,8	4,9	923,2	548,9	935,7	550,7	894,1	563,5	0,019	17
11	2	2o	0,017	0,024	0,037	0,015	0,90	0,25	0,00	-	1,02	0,46	0,98	0,47	0,09	-	35,1	10,2	964,6	546,1	1009,3	520,8	919,4	549,9	0,014	19
28	2	2o	0,016	0,023	0,034	0,008	0,93	0,18	0,00	-	0,98	0,15	0,96	0,17	0,09	-	12,0	1,6	943,4	503,0	968,4	523,1	924,6	553,6	0,009	21
10	2	2o	0,020	0,025	0,035	0,012	0,87	0,27	0,00	-	1,13	0,93	1,12	1,04	0,20	-	23,5	3,7	884,9	625,8	982,1	661,4	930,0	687,1	0,045	7
9	2	2o	0,016	0,017	0,034	0,009	0,92	0,18	0,00	-	0,96	0,12	0,94	0,15	0,10	-	22,2	12,4	896,6	516,4	895,9	546,6	866,0	574,0	0,002	33
35	4	3r1o	0,012	0,018	0,061	0,208	0,93	0,20	0,01	-	0,99	0,11	0,89	0,26	0,10	-	153,2	59,2	1494,4	1269,3	1481,6	1258,4	1214,0	656,7	0,006	26
40	4	2r2o	0,010	0,009	0,043	0,032	0,95	0,18	0,00	-	1,00	0,18	0,93	0,31	0,09	-	121,6	64,8	1358,5	701,9	1355,5	694,6	1232,0	617,2	0,010	20
37	4	2r2o	0,008	0,006	0,040	0,019	0,94	0,17	0,00	-	1,01	0,10	0,99	0,42	0,15	-	87,1	63,0	1336,9	698,9	1335,8	662,5	1227,4	565,8	0,001	34
36	4	2r2o	0,010	0,007	0,036	0,014	0,93	0,18	0,00	-	1,11	0,73	1,06	0,83	0,16	-	83,1	64,6	1130,1	636,9	1159,4	631,6	1101,6	662,3	0,031	10
39	4	1r3o	0,009	0,007	0,035	0,011	0,92	0,20	0,00	-	1,00	0,10	0,97	0,22	0,15	-	79,9	66,1	1114,9	483,7	1117,9	487,6	1067,7	458,7	0,005	27
38	4	1r3o	0,010	0,009	0,043	0,034	0,94	0,16	0,00	-	1,06	0,35	1,00	0,40	0,15	-	50,9	7,2	1189,4	685,9	1202,3	637,2	1095,7	579,9	0,028	12
34	4	4o	0,017	0,022	0,036	0,014	0,89	0,25	0,00	-	1,16	0,71	1,11	0,70	0,23	-	40,5	12,5	893,5	563,8	1013,8	550,4	921,6	568,5	0,084	4
41	8	3r5o	0,010	0,011	0,036	0,009	0,92	0,23	0,00	-	1,02	0,26	0,98	0,34	0,15	-	59,9	29,2	1166,0	606,8	1185,1	615,5	1129,3	605,8	0,026	13
42	16	6r10o	0,010	0,010	0,037	0,017	0,92	0,22	0,00	-	1,00	0,11	0,95	0,24	0,15	-	80,5	60,4	1190,5	636,5	1198,8	654,9	1109,9	579,3	0,007	24

Tabella2.6: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 750 I_C 10. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,006	0,003	0,018	0,011	0,79	0,11	0,00	-	1,01	0,14	0,97	0,32	0,07	-	44,6	49,2	7247,0	1972,3	7276,0	1900,8	6809,8	1887,6	0,017	20
7	1	controllo	0,006	0,004	0,018	0,011	0,78	0,12	0,00	-	1,00	0,04	0,95	0,16	0,08	-	19,0	5,6	7204,8	2060,4	7202,8	2081,8	6723,7	2088,1	0,001	33
6	1	controllo	0,007	0,006	0,022	0,020	0,82	0,06	0,01	-	1,00	0,07	0,95	0,30	0,08	-	81,5	58,0	7812,3	2916,0	7717,4	2660,9	6982,7	1847,9	0,001	32
5	1	1o	0,005	0,002	0,013	0,004	0,78	0,25	0,01	-	1,16	0,23	1,48	1,29	0,51	-	40,7	46,0	4131,0	1110,0	4714,6	1193,9	5558,7	2050,4	0,001	30
4	1	1o	0,006	0,003	0,012	0,004	0,67	0,27	0,01	-	1,20	0,31	1,51	0,94	0,44	-	51,3	76,0	3281,2	1092,9	3861,6	1265,5	4714,7	2277,7	0,000	38
3	1	1o	0,005	0,004	0,012	0,003	0,70	0,25	0,00	-	1,24	0,29	1,59	0,81	0,53	-	36,7	31,2	3047,1	984,7	3705,8	1195,6	4592,6	1909,4	0,001	36
2	1	1o	0,005	0,004	0,012	0,004	0,70	0,26	0,01	-	1,27	0,34	1,79	1,27	0,55	-	57,6	56,5	2997,9	1050,5	3663,1	1191,8	4848,7	2288,3	0,000	40
1	1	1o	0,005	0,003	0,013	0,004	0,73	0,26	0,03	-	1,28	0,34	1,99	1,63	0,56	-	68,3	81,7	2884,4	1028,9	3605,8	1250,9	5099,8	2890,5	0,019	17
27	2	1r1o	0,005	0,002	0,014	0,006	0,88	0,19	0,00	-	1,05	0,08	1,35	0,83	0,47	-	174,6	14,2	5520,5	1941,2	5736,3	1885,4	6573,1	1907,8	0,003	29
24	2	1r1o	0,004	0,002	0,012	0,004	0,86	0,21	0,02	-	1,10	0,16	1,78	1,42	0,54	-	182,1	13,0	4579,3	1916,0	4896,8	1827,1	6485,6	2080,1	0,001	34
21	2	1r1o	0,005	0,003	0,013	0,004	0,78	0,32	0,02	-	1,10	0,17	1,76	1,21	0,55	-	196,5	20,5	3993,5	1712,6	4289,3	1656,8	5832,2	2054,7	0,000	41
18	2	1r1o	0,004	0,003	0,013	0,005	0,81	0,27	0,07	-	1,11	0,15	2,22	1,76	0,64	-	204,7	17,0	3674,7	1754,1	3985,9	1748,2	6077,3	2056,7	0,000	39
15	2	1r1o	0,004	0,003	0,013	0,005	0,78	0,29	0,10	-	1,09	0,10	2,28	2,04	0,68	-	215,2	17,1	3660,3	1786,3	3906,5	1800,6	6000,3	2187,1	0,000	42
26	2	1r1o	0,005	0,002	0,014	0,005	0,90	0,19	0,02	-	1,65	1,92	1,67	1,90	0,48	-	65,7	9,4	5523,4	2259,2	6910,8	2157,2	6893,3	2165,5	0,024	13
23	2	1r1o	0,005	0,003	0,015	0,009	0,89	0,20	0,01	-	1,85	1,68	1,98	1,93	0,61	-	82,0	25,0	4524,0	2094,1	6109,3	1857,4	6287,1	1944,1	0,035	10
20	2	1r1o	0,004	0,003	0,014	0,008	0,88	0,25	0,07	-	3,16	3,09	3,10	2,79	0,74	-	87,0	12,3	3462,6	2269,1	6299,8	1983,0	6342,8	1898,4	0,078	3
17	2	1r1o	0,003	0,003	0,013	0,004	0,86	0,23	0,11	-	4,06	4,35	4,52	4,33	0,79	-	99,0	17,6	2563,9	1759,0	5646,5	1893,5	6387,9	2097,2	0,056	6
14	2	1r1o	0,004	0,002	0,014	0,007	0,89	0,20	0,13	-	3,29	3,21	4,05	3,63	0,86	-	109,1	16,2	2709,5	1735,3	5426,2	1716,0	6557,4	2042,1	0,035	9
25	2	1r1o	0,006	0,004	0,019	0,019	0,84	0,13	0,00	-	1,08	0,25	1,04	0,36	0,19	-	20,5	17,2	6687,4	2374,7	7037,7	2265,0	6573,3	2063,7	0,001	37
22	2	1r1o	0,005	0,002	0,015	0,007	0,89	0,12	0,00	-	2,17	3,62	2,24	3,79	0,42	-	29,0	20,9	5222,3	2235,8	6574,2	1764,8	6506,3	1877,0	0,026	11
19	2	1r1o	0,005	0,003	0,016	0,009	0,90	0,19	0,05	-	4,37	9,71	4,30	9,51	0,67	-	34,2	5,9	4139,4	2661,3	6398,8	2185,6	6248,8	2149,9	0,073	5
16	2	1r1o	0,004	0,003	0,014	0,007	0,92	0,20	0,06	-	5,35	6,99	5,38	7,13	0,79	-	44,0	5,3	3340,8	2489,7	6789,4	1955,2	6755,9	2198,4	0,112	1
13	2	1r1o	0,004	0,003	0,014	0,005	0,89	0,25	0,17	-	6,27	8,45	6,10	7,98	0,76	-	54,3	9,2	2964,3	2313,7	6391,5	2199,7	6467,5	2249,8	0,056	8
33	2	2o	0,005	0,003	0,012	0,003	0,71	0,28	0,00	-	1,40	1,72	1,54	2,00	0,40	-	30,5	24,9	3800,3	1257,6	4501,7	1419,5	4914,0	1863,3	0,010	26
32	2	2o	0,005	0,003	0,012	0,004	0,78	0,26	0,03	-	2,48	3,41	2,89	3,71	0,65	-	38,5	26,9	2979,6	1616,2	4401,6	1544,6	5149,3	2191,6	0,025	12
30	2	2o	0,005	0,004	0,012	0,004	0,74	0,29	0,06	-	3,93	6,64	4,38	6,81	0,74	-	45,6	14,0	2550,3	1561,5	4282,7	1400,0	4987,5	1802,2	0,024	15
12	2	2o	0,004	0,003	0,012	0,003	0,72	0,31	0,10	-	4,40	6,11	4,98	6,24	0,74	-	63,1	51,4	2137,8	1488,8	4439,6	1630,3	5256,6	2176,8	0,023	16
31	2	2o	0,006	0,004	0,012	0,004	0,71	0,27	0,00	-	1,30	0,50	1,60	1,07	0,44	-	39,2	32,1	3152,9	1101,4	3890,9	1312,2	4632,0	2215,7	0,001	31
29	2	2o	0,005	0,003	0,012	0,003	0,70	0,29	0,01	-	2,29	2,86	2,74	3,29	0,66	-	41,9	33,0	2730,2	1321,1	4020,1	1463,3	4855,2	2157,4	0,013	24
11	2	2o	0,005	0,004	0,012	0,003	0,69	0,28	0,04	-	3,26	5,41	3,94	5,84	0,71	-	56,5	38,3	2224,0	1285,8	4019,6	1664,3	4949,8	2429,9	0,017	19
28	2	2o	0,006	0,003	0,012	0,003	0,73	0,27	0,01	-	1,38	1,05	1,87	2,08	0,61	-	46,2	65,9	3067,6	1115,0	3831,5	1291,1	4834,8	2422,2	0,003	28
10	2	2o	0,005	0,004	0,012	0,003	0,71	0,28	0,01	-	2,19	2,48	2,64	2,87	0,71	-	42,6	28,3	2495,3	1137,7	3879,4	1361,8	4702,1	2145,9	0,017	21
9	2	2o	0,006	0,003	0,012	0,004	0,65	0,26	0,01	-	1,70	2,26	2,30	3,63	0,52	-	47,8	47,5	2782,2	1009,9	3625,5	1254,4	4603,7	2326,6	0,001	35
35	4	3r1o	0,005	0,003	0,015	0,007	0,92	0,20	0,00	-	1,61	2,30	1,56	2,07	0,30	-	110,2	76,2	5960,2	2416,7	6683,4	2016,5	6677,8	2266,5	0,005	27
40	4	2r2o	0,005	0,002	0,014	0,005	0,92	0,20	0,01	-	1,48	1,73	1,56	1,75	0,42	-	86,1	46,9	5577,2	2128,8	6443,2	1887,8	6724,6	2003,8	0,024	14
37	4	2r2o	0,005	0,003	0,014	0,006	0,92	0,22	0,01	-	2,07	3,32	2,17	3,28	0,46	-	70,9	52,0	5064,5	2339,2	6177,7	1819,8	6444,8	1980,0	0,016	22
36	4	2r2o	0,004	0,003	0,014	0,006	0,92	0,21	0,03	-	2,43	3,38	2,63	3,52	0,56	-	66,9	47,0	4367,2	2051,8	6022,3	1672,6	6439,4	2013,8	0,011	25
39	4	1r3o	0,005	0,003	0,013	0,006	0,80	0,30	0,11	-	3,07	6,87	4,15	7,70	0,61	-	112,3	73,9	3344,2	1928,3	4536,4	1881,7	5913,9	2278,2	0,018	18
38	4	1r3o	0,004	0,002	0,013	0,004	0,87	0,25	0,08	-	5,14	8,56	5,29	8,57	0,75	-	52,5	11,6	3325,6	2308,1	5909,6	1964,0	6200,2	2183,5	0,075	4
34	4	4o	0,005	0,003	0,012	0,004	0,76	0,30	0,04	-	4,20	6,64	4,65	7,04	0,73	-	51,5	20,2	2394,7	1513,3	4714,7	1805,4	5315,1	2329,9	0,056	7
41	8	3r5o	0,004	0,002	0,014	0,006	0,91	0,22	0,05	-	3,28	5,39	3,63	5,78	0,72	-	71,5	44,5	3879,1	2284,2	6041,7	1789,3	6564,7	2115,1	0,107	2
42	16	6r10o	0,004	0,002	0,013	0,005	0,91	0,20	0,02	-	2,28	3,94	2,51	4,21	0,51	-	75,9	50,3	4754,6	2341,8	5942,3	1716,2	6364,3	1690,2	0,013	23

Tabella2.7: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 750 I_C 100. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,007	0,004	0,011	0,008	0,94	0,12	0,00	-	1,00	0,02	0,98	0,13	0,06	-	115,7	64,5	76069,9	12514,0	76324,5	12355,2	73780,3	12668,1	0,000	35
7	1	controllo	0,007	0,003	0,011	0,010	0,95	0,09	0,00	-	1,00	0,00	0,97	0,09	0,01	-	185,0	NA	77174,9	11554,3	77302,3	11512,2	74638,2	10126,8	0,000	34
6	1	controllo	0,007	0,004	0,010	0,007	0,95	0,10	0,01	-	1,00	0,00	0,97	0,13	0,04	-	369,0	382,9	74834,9	10890,0	74906,8	10838,9	72143,0	12327,5	0,004	28
5	1	1o	0,001	0,000	0,003	0,001	0,69	0,20	0,99	-	1,74	1,70	16,74	8,36	1,00	-	299,5	135,2	4652,8	1558,6	7016,1	3031,4	69104,7	18301,4	0,000	41
4	1	1o	0,001	0,001	0,004	0,002	0,74	0,20	0,98	-	1,72	0,93	23,08	9,28	1,00	-	274,0	122,5	3269,2	1039,5	5202,1	1881,9	69269,9	19674,7	0,000	33
3	1	1o	0,001	0,001	0,004	0,002	0,75	0,19	0,98	-	1,77	1,00	25,70	10,00	1,00	-	258,4	113,1	2798,6	919,1	4620,4	2236,9	65903,4	18248,6	0,001	31
2	1	1o	0,001	0,001	0,004	0,001	0,75	0,16	1,00	-	1,64	1,34	29,81	13,82	1,00	-	259,9	81,1	2462,9	693,4	3741,8	1688,3	67168,2	21091,3	0,000	38
1	1	1o	0,001	0,001	0,004	0,002	0,72	0,16	0,99	-	1,49	0,43	29,31	13,16	1,00	-	285,7	104,8	2400,1	800,6	3379,4	780,2	64160,6	20891,6	0,001	30
27	2	1r1o	0,001	0,002	0,005	0,003	0,89	0,11	0,47	-	1,54	0,73	13,50	8,49	0,97	-	167,9	37,1	12523,7	16797,8	16360,4	17904,6	72588,2	12197,2	0,000	36
24	2	1r1o	0,001	0,001	0,003	0,002	0,86	0,14	0,77	-	1,34	0,58	22,35	10,28	0,99	-	171,5	20,6	5268,7	8315,5	6787,2	9616,8	72910,9	13820,2	0,000	37
21	2	1r1o	0,001	0,000	0,003	0,002	0,87	0,12	0,96	-	1,19	0,13	33,17	12,70	1,00	-	177,8	15,4	2534,9	1038,4	3054,0	1400,9	73701,4	13790,4	0,000	40
18	2	1r1o	0,001	0,000	0,003	0,001	0,82	0,19	0,98	-	1,17	0,11	34,34	14,20	1,00	-	188,2	15,8	2313,2	780,1	2699,2	947,2	70612,4	14305,5	0,000	32
15	2	1r1o	0,000	0,000	0,003	0,001	0,85	0,11	1,00	-	1,14	0,11	37,24	10,67	1,00	-	195,3	15,6	2024,6	613,1	2323,4	782,0	70306,7	11594,5	0,000	39
26	2	1r1o	0,002	0,002	0,007	0,008	0,88	0,25	0,25	-	20,01	23,64	20,47	23,31	0,87	-	70,0	34,6	21761,0	23868,0	63498,9	15056,4	69350,4	14027,7	0,027	15
23	2	1r1o	0,001	0,000	0,004	0,003	0,92	0,19	0,78	-	47,43	23,21	47,39	22,35	1,00	-	72,2	10,9	2578,3	5238,8	69650,9	13856,2	70771,5	13971,7	0,037	12
20	2	1r1o	0,000	0,000	0,004	0,002	0,94	0,14	0,90	-	61,03	27,13	61,00	26,80	1,00	-	80,9	6,6	1585,4	1919,6	71101,1	13695,5	71966,5	13436,8	0,041	10
17	2	1r1o	0,000	0,000	0,004	0,002	0,94	0,13	1,00	-	63,93	25,97	65,75	28,52	1,00	-	89,9	7,9	1182,3	352,0	68787,0	14240,4	70036,5	12733,1	0,102	1
14	2	1r1o	0,000	0,000	0,004	0,002	0,95	0,06	0,99	-	63,67	21,43	73,35	25,57	1,00	-	100,4	7,3	1043,6	319,1	61148,9	10899,1	70178,5	12379,3	0,041	11
25	2	1r1o	0,005	0,003	0,008	0,004	0,87	0,22	0,11	-	1,19	0,52	1,47	0,89	0,56	-	415,3	2456,0	56722,1	16141,7	63563,8	15286,9	73112,5	17654,3	0,000	42
22	2	1r1o	0,005	0,002	0,007	0,003	0,91	0,09	0,58	-	75,44	85,61	75,91	86,22	0,95	-	25,7	42,5	11915,1	18207,5	70345,3	13821,5	73144,3	12892,7	0,072	4
19	2	1r1o	0,004	0,002	0,007	0,003	0,91	0,16	0,91	-	138,48	60,29	136,22	58,70	1,00	-	31,1	7,5	1285,2	5035,2	72178,2	13666,2	72182,1	13030,3	0,047	9
16	2	1r1o	0,003	0,002	0,006	0,003	0,91	0,17	0,96	-	142,33	52,63	140,78	51,21	1,00	-	40,1	4,3	580,0	256,9	72763,9	13578,4	72460,5	14213,7	0,090	2
13	2	1r1o	0,001	0,001	0,005	0,003	0,92	0,18	1,00	-	139,00	57,58	138,26	56,12	1,00	-	50,6	5,9	611,5	373,4	73416,5	13194,0	73198,4	13030,9	0,087	3
33	2	2o	0,001	0,001	0,005	0,003	0,79	0,24	0,98	-	20,96	46,38	36,88	45,69	0,99	-	178,5	122,3	3032,9	1727,6	23535,5	24566,1	67600,9	20040,3	0,005	27
32	2	2o	0,001	0,001	0,004	0,002	0,79	0,21	1,00	-	62,77	89,96	82,16	90,27	1,00	-	146,6	135,0	1969,0	1703,6	34059,6	27424,7	68718,7	19329,9	0,015	21
30	2	2o	0,001	0,001	0,004	0,002	0,81	0,20	0,99	-	80,54	83,57	98,08	78,87	1,00	-	124,3	106,7	1345,6	1251,2	41738,0	27809,4	67820,4	18015,9	0,023	16
12	2	2o	0,001	0,001	0,003	0,001	0,77	0,21	0,99	-	82,32	76,66	100,64	78,34	1,00	-	99,5	57,5	993,7	790,0	46013,2	23437,2	65369,0	17409,3	0,016	20
31	2	2o	0,001	0,001	0,004	0,002	0,75	0,21	0,99	-	16,76	47,63	38,39	47,20	1,00	-	216,5	103,8	2586,9	1267,9	13470,7	17901,6	64128,2	17640,2	0,002	29
29	2	2o	0,001	0,001	0,004	0,002	0,81	0,19	1,00	-	55,40	74,92	79,26	70,00	1,00	-	151,9	145,7	1490,4	1179,4	30791,7	25871,2	66990,0	19183,3	0,016	19
11	2	2o	0,001	0,001	0,003	0,001	0,77	0,21	1,00	-	59,57	73,71	82,92	71,99	1,00	-	146,9	96,3	1412,6	1090,1	32583,1	26804,1	65766,9	16609,2	0,014	22
28	2	2o	0,001	0,001	0,004	0,002	0,75	0,22	1,00	-	19,69	40,30	43,96	40,69	1,00	-	217,0	153,0	2280,8	1866,6	16187,5	21870,1	62535,8	17341,5	0,006	26
10	2	2o	0,001	0,001	0,004	0,001	0,76	0,20	1,00	-	47,64	88,62	75,12	85,71	1,00	-	172,8	106,5	1616,1	1043,2	22410,3	24449,1	64364,6	18317,2	0,008	24
9	2	2o	0,001	0,001	0,004	0,002	0,79	0,18	0,98	-	16,53	36,08	46,63	38,03	1,00	-	219,7	118,2	1998,3	1094,4	12794,1	18896,8	66203,5	19769,8	0,006	25
35	4	3r1o	0,004	0,004	0,008	0,005	0,94	0,18	0,22	-	40,87	62,88	41,50	62,19	0,69	-	84,1	124,3	34033,3	30340,6	65021,9	16170,0	69907,2	12748,0	0,016	18
40	4	2r2o	0,002	0,003	0,006	0,004	0,89	0,22	0,32	-	19,25	25,39	22,14	25,74	0,91	-	100,0	64,6	16140,3	21732,7	55332,6	20555,7	69263,7	12050,4	0,011	23
37	4	2r2o	0,002	0,002	0,005	0,004	0,93	0,21	0,40	-	74,41	68,93	75,71	69,57	0,87	-	65,5	79,0	14978,9	23890,6	66352,1	15423,1	70729,4	13012,6	0,031	14
36	4	2r2o	0,001	0,001	0,005	0,003	0,92	0,16	0,94	-	123,03	63,55	125,73	60,78	1,00	-	59,1	52,2	2052,4	7124,5	67006,2	15554,8	71594,0	12083,1	0,058	5
39	4	1r3o	0,000	0,000	0,003	0,001	0,84	0,18	1,00	-	55,46	69,83	73,69	67,29	1,00	-	128,5	74,7	1774,0	1379,6	35953,6	27780,8	70763,9	12156,1	0,020	17
38	4	1r3o	0,000	0,000	0,003	0,001	0,80	0,23	0,97	-	161,61	62,62	161,85	63,54	1,00	-	52,3	9,9	492,4	262,8	68012,7	11881,0	68305,3	11973,4	0,058	6
34	4	4o	0,001	0,000	0,003	0,001	0,83	0,17	1,00	-	86,03	65,67	101,52	62,45	1,00	-	94,5	78,0	978,0	942,7	48665,0	21779,9	65883,8	14510,7	0,034	13
41	8	3r5o	0,001	0,001	0,004	0,002	0,89	0,21	0,85	-	99,77	64,34	100,10	61,66	1,00	-	63,6	24,2	2507,5	6708,3	66275,7	18375,0	71120,8	15233,3	0,055	7
42	16	6r10o	0,001	0,001	0,004	0,002	0,93	0,15	0,84	-	88,82	70,56	92,64	69,69	1,00	-	76,3	55,5	2103,7	4283,6	64210,0	20723,5	72663,1	12580,5	0,055	8

Tabella2.8: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 1500 I_C 2. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,005	0,004	0,023	0,008	0,97	0,08	0,00	-	0,96	0,07	0,90	0,15	0,04	-	20,3	11,7	2970,8	1103,6	2856,3	1074,2	2653,0	1064,4	0,002	40
7	1	controllo	0,006	0,012	0,023	0,009	0,97	0,11	0,00	-	0,98	0,05	0,91	0,15	0,01	-	38,0	-	3079,2	1163,3	3035,0	1094,4	2772,7	1089,7	0,000	42
6	1	controllo	0,005	0,005	0,024	0,008	0,99	0,04	0,00	-	0,96	0,09	0,90	0,17	0,05	-	18,4	8,0	2883,4	1088,9	2778,4	1056,2	2571,6	978,7	0,003	38
5	1	1o	0,006	0,004	0,020	0,005	0,97	0,10	0,00	-	1,01	0,08	1,01	0,23	0,10	-	26,8	18,4	2311,2	831,1	2340,0	871,1	2319,9	922,8	0,003	39
4	1	1o	0,006	0,006	0,020	0,005	0,95	0,11	0,00	-	1,01	0,06	1,01	0,12	0,14	-	28,5	26,9	2403,2	990,7	2445,9	1039,4	2451,0	1095,4	0,007	32
3	1	1o	0,009	0,008	0,020	0,005	0,93	0,14	0,00	-	1,00	0,12	0,98	0,24	0,07	-	33,6	20,5	1931,0	812,0	1951,0	886,6	1937,7	989,4	0,006	34
2	1	1o	0,006	0,006	0,019	0,005	0,94	0,13	0,00	-	1,01	0,08	1,05	0,47	0,14	-	45,9	63,7	2162,7	787,8	2202,2	871,0	2271,2	1085,9	0,004	37
1	1	1o	0,006	0,005	0,019	0,004	0,94	0,12	0,00	-	1,01	0,06	1,00	0,11	0,09	-	27,8	13,6	2175,2	887,6	2198,5	921,4	2183,8	965,1	0,008	31
27	2	1r1o	0,005	0,004	0,022	0,007	0,93	0,20	0,00	-	0,99	0,02	0,94	0,22	0,09	-	168,4	5,1	2668,4	1032,0	2650,9	1031,7	2458,9	947,5	0,030	14
24	2	1r1o	0,004	0,003	0,021	0,006	0,98	0,11	0,00	-	1,00	0,03	0,99	0,24	0,13	-	191,8	18,3	2742,2	841,9	2743,7	839,9	2681,5	950,9	0,010	29
21	2	1r1o	0,005	0,004	0,022	0,007	0,97	0,13	0,00	-	1,01	0,04	1,03	0,34	0,21	-	197,8	17,0	2766,6	1026,2	2776,2	1015,6	2725,8	966,7	0,002	41
18	2	1r1o	0,005	0,004	0,021	0,009	0,96	0,14	0,00	-	1,00	0,06	1,08	0,71	0,15	-	208,2	17,4	2658,2	977,7	2660,0	963,7	2644,3	1035,0	0,037	9
15	2	1r1o	0,005	0,004	0,022	0,009	0,95	0,17	0,00	-	1,00	0,04	0,96	0,25	0,09	-	216,6	15,0	2576,8	904,1	2580,4	920,0	2472,3	1068,9	0,019	20
26	2	1r1o	0,005	0,005	0,025	0,013	0,98	0,09	0,00	-	1,00	0,12	0,93	0,24	0,12	-	72,4	15,5	2985,8	1263,5	2970,8	1236,6	2677,9	1051,5	0,020	19
23	2	1r1o	0,006	0,005	0,024	0,010	0,95	0,14	0,00	-	0,99	0,12	0,94	0,26	0,12	-	85,4	24,3	2719,3	1139,7	2698,3	1160,8	2521,9	1182,7	0,005	36
20	2	1r1o	0,006	0,006	0,021	0,006	0,93	0,20	0,00	-	1,02	0,13	0,98	0,21	0,15	-	104,2	46,3	2513,2	927,0	2566,5	941,0	2455,5	989,6	0,027	16
17	2	1r1o	0,005	0,004	0,025	0,016	0,95	0,16	0,00	-	1,11	0,38	1,08	0,50	0,23	-	101,6	14,0	2688,5	1345,7	2820,2	1265,0	2633,1	1153,7	0,081	2
14	2	1r1o	0,006	0,005	0,021	0,009	0,94	0,19	0,00	-	1,06	0,27	1,05	0,49	0,21	-	109,8	12,0	2445,8	1096,6	2526,6	1091,9	2424,0	1050,1	0,047	5
25	2	1r1o	0,005	0,004	0,024	0,009	0,97	0,09	0,00	-	0,97	0,08	0,90	0,17	0,04	-	17,8	4,7	2824,9	1050,1	2738,3	1020,0	2508,9	960,1	0,008	30
22	2	1r1o	0,006	0,006	0,024	0,009	0,97	0,09	0,00	-	0,98	0,13	0,92	0,20	0,10	-	27,1	6,5	2695,9	964,2	2646,5	953,8	2451,5	963,3	0,005	35
19	2	1r1o	0,006	0,005	0,023	0,008	0,91	0,17	0,00	-	1,02	0,16	0,95	0,22	0,16	-	38,3	7,9	2514,7	1079,2	2542,8	1091,5	2326,8	1029,6	0,014	25
16	2	1r1o	0,005	0,004	0,025	0,020	0,99	0,08	0,00	-	1,07	0,53	1,02	0,57	0,14	-	45,7	9,9	2660,8	1047,8	2714,7	1000,6	2538,6	949,7	0,077	3
13	2	1r1o	0,005	0,004	0,021	0,006	0,95	0,16	0,00	-	1,10	0,39	1,07	0,46	0,21	-	57,4	10,2	2536,4	1014,3	2676,3	1032,5	2580,0	1055,5	0,072	4
33	2	2o	0,005	0,006	0,020	0,005	0,96	0,13	0,01	-	1,02	0,11	1,02	0,18	0,21	-	24,5	14,7	2357,6	801,9	2420,0	870,1	2421,0	958,6	0,015	22
32	2	2o	0,007	0,006	0,020	0,005	0,92	0,15	0,00	-	1,00	0,10	0,99	0,18	0,13	-	36,2	17,5	2265,8	1042,4	2281,3	1074,3	2262,0	1131,8	0,006	33
30	2	2o	0,006	0,005	0,022	0,012	0,95	0,13	0,00	-	1,01	0,15	0,96	0,24	0,13	-	44,6	19,6	2370,3	1020,8	2363,0	951,4	2217,1	917,4	0,025	18
12	2	2o	0,006	0,005	0,021	0,006	0,94	0,16	0,00	-	1,35	2,09	1,35	2,22	0,21	-	49,8	10,2	2256,6	955,2	2464,8	991,2	2432,7	1019,9	0,084	1
31	2	2o	0,007	0,007	0,020	0,005	0,94	0,12	0,00	-	1,04	0,32	1,04	0,45	0,18	-	24,3	14,7	2153,5	873,6	2221,1	945,6	2219,9	1030,6	0,027	17
29	2	2o	0,006	0,005	0,021	0,009	0,96	0,11	0,00	-	1,01	0,15	0,98	0,21	0,15	-	28,1	11,3	2344,4	866,9	2365,1	892,0	2282,0	925,4	0,035	10
11	2	2o	0,006	0,005	0,020	0,005	0,94	0,15	0,00	-	1,05	0,19	1,03	0,24	0,11	-	41,3	19,9	2094,3	762,3	2166,1	797,3	2132,0	831,6	0,030	15
28	2	2o	0,010	0,015	0,020	0,006	0,91	0,19	0,00	-	1,00	0,14	0,98	0,19	0,08	-	17,1	7,3	1973,6	865,1	2022,8	933,0	1956,7	995,3	0,011	26
10	2	2o	0,006	0,007	0,020	0,008	0,94	0,16	0,00	-	1,06	0,21	1,05	0,26	0,22	-	27,0	6,5	2159,6	836,5	2260,9	889,8	2242,3	920,5	0,030	13
9	2	2o	0,007	0,008	0,020	0,004	0,93	0,16	0,00	-	1,01	0,14	1,00	0,21	0,10	-	27,3	15,4	2092,7	886,2	2125,5	958,7	2120,7	1028,1	0,031	12
35	4	3r1o	0,005	0,003	0,023	0,010	0,98	0,12	0,00	-	1,02	0,06	0,96	0,35	0,10	-	117,3	31,9	2913,6	1185,1	2952,7	1177,0	2681,9	1013,5	0,014	24
40	4	2r2o	0,005	0,004	0,021	0,006	0,96	0,13	0,00	-	1,01	0,13	0,97	0,24	0,11	-	115,0	58,5	2611,9	1056,1	2630,2	1062,2	2496,7	1034,3	0,035	11
37	4	2r2o	0,005	0,004	0,022	0,009	0,96	0,16	0,00	-	1,07	0,55	1,03	0,67	0,12	-	79,6	60,1	2694,7	1047,1	2747,3	1020,4	2578,4	976,6	0,017	21
36	4	2r2o	0,005	0,004	0,021	0,007	0,96	0,14	0,00	-	1,02	0,08	1,01	0,33	0,14	-	126,5	81,5	2604,3	946,9	2636,7	954,4	2580,4	1014,9	0,010	28
39	4	1r3o	0,005	0,006	0,021	0,011	0,93	0,19	0,00	-	1,01	0,09	1,04	0,45	0,17	-	100,1	78,9	2505,1	1050,7	2533,0	1040,5	2504,2	1008,8	0,010	27
38	4	1r3o	0,005	0,006	0,023	0,007	0,97	0,13	0,00	-	1,07	0,30	1,01	0,33	0,19	-	50,4	6,1	2773,5	1091,6	2868,2	1046,8	2681,1	971,8	0,039	7
34	4	4o	0,006	0,007	0,022	0,009	0,97	0,11	0,00	-	1,10	0,39	1,08	0,48	0,25	-	44,8	11,5	2285,9	845,6	2430,9	883,0	2380,4	945,1	0,045	6
41	8	3r5o	0,006	0,005	0,023	0,011	0,92	0,22	0,00	-	1,05	0,18	0,98	0,26	0,15	-	84,4	58,7	2620,3	1164,2	2684,3	1135,9	2471,1	1099,4	0,037	8
42	16	6r10o	0,005	0,004	0,024	0,010	0,97	0,12	0,00	-	1,02	0,11	0,96	0,25	0,18	-	90,4	70,9	2860,2	987,7	2892,8	966,0	2684,0	972,9	0,015	23

Tabella2.9: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 1500 I_C 10. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,003	0,002	0,009	0,004	0,88	0,08	0,00	-	1,00	0,02	0,94	0,13	0,04	-	64,0	86,7	14950,4	3190,2	14953,3	3225,4	13958,2	3291,8	0,016	17
7	1	controllo	0,003	0,003	0,011	0,010	0,88	0,10	0,01	-	1,01	0,03	0,94	0,19	0,10	-	47,7	22,8	15596,2	3862,6	15649,1	3775,7	14355,7	3561,8	0,014	18
6	1	controllo	0,003	0,001	0,009	0,005	0,88	0,06	0,00	-	1,01	0,02	0,98	0,11	0,07	-	70,7	103,7	14874,6	3150,9	14941,6	3052,4	14460,0	3048,6	0,004	30
5	1	1o	0,003	0,001	0,007	0,002	0,73	0,32	0,02	-	1,15	0,31	1,42	0,84	0,53	-	58,4	75,4	8959,8	2726,7	9991,2	2667,0	11631,4	3790,6	0,002	35
4	1	1o	0,003	0,002	0,007	0,002	0,68	0,34	0,02	-	1,29	0,41	1,93	1,49	0,56	-	75,6	93,1	6943,6	2275,4	8410,5	2038,7	11261,1	4196,6	0,002	34
3	1	1o	0,002	0,002	0,007	0,002	0,72	0,32	0,05	-	1,35	0,48	2,52	2,13	0,70	-	89,0	77,4	6408,1	2496,8	8001,6	2415,9	12463,8	4649,0	0,002	36
2	1	1o	0,003	0,002	0,008	0,002	0,74	0,28	0,06	-	1,53	0,57	2,61	1,83	0,79	-	86,2	95,8	5534,4	2198,7	7679,1	2108,8	12050,4	5117,8	0,008	25
1	1	1o	0,003	0,002	0,008	0,002	0,71	0,29	0,05	-	1,46	0,56	2,66	2,12	0,79	-	91,2	97,6	5448,8	2042,0	7339,1	2030,7	11606,3	4757,0	0,023	12
27	2	1r1o	0,003	0,002	0,008	0,004	0,92	0,14	0,00	-	1,02	0,11	1,11	0,66	0,16	-	203,1	62,0	13824,0	3625,7	13983,1	3478,1	14208,9	3448,3	0,003	33
24	2	1r1o	0,003	0,001	0,007	0,003	0,89	0,21	0,00	-	1,02	0,04	1,16	0,55	0,21	-	199,8	61,6	12436,5	3006,5	12661,8	3001,4	13506,8	2777,1	0,002	40
21	2	1r1o	0,003	0,001	0,008	0,003	0,88	0,22	0,03	-	1,05	0,09	1,67	1,36	0,51	-	249,5	223,4	10285,9	3514,4	10661,5	3345,0	14123,1	5229,6	0,002	38
18	2	1r1o	0,003	0,002	0,008	0,004	0,92	0,15	0,02	-	1,11	0,19	2,01	1,61	0,58	-	205,3	44,7	9209,7	3941,7	9840,2	3669,1	13392,4	3304,0	0,001	41
15	2	1r1o	0,002	0,001	0,007	0,002	0,86	0,26	0,02	-	1,15	0,24	2,14	1,45	0,66	-	211,0	32,0	8203,8	4006,4	9004,0	3810,7	13200,9	3049,8	0,005	27
26	2	1r1o	0,003	0,001	0,009	0,004	0,95	0,07	0,00	-	1,06	0,15	1,06	0,27	0,19	-	98,8	56,7	13534,8	3406,9	14090,0	3187,7	13974,0	3257,9	0,003	31
23	2	1r1o	0,003	0,002	0,009	0,005	0,92	0,19	0,02	-	1,38	1,55	1,47	1,97	0,27	-	135,3	253,6	12867,2	3934,3	14093,4	2981,9	13987,5	3507,5	0,021	13
20	2	1r1o	0,002	0,001	0,008	0,003	0,93	0,20	0,01	-	1,93	2,35	2,07	2,45	0,53	-	96,5	22,0	9966,0	4038,1	12711,0	2662,8	13428,9	3155,2	0,100	2
17	2	1r1o	0,002	0,001	0,008	0,003	0,91	0,20	0,02	-	2,27	2,70	2,54	2,78	0,59	-	107,6	54,9	9073,1	4854,6	12488,1	3337,2	13718,7	3507,3	0,076	5
14	2	1r1o	0,002	0,001	0,008	0,003	0,93	0,17	0,09	-	2,58	2,94	3,14	3,55	0,73	-	108,8	19,5	8147,7	4612,1	12212,3	3146,8	13907,8	3209,5	0,094	3
25	2	1r1o	0,003	0,002	0,009	0,005	0,87	0,16	0,00	-	1,02	0,04	0,97	0,19	0,10	-	35,6	22,1	14581,0	4842,6	14815,9	4800,7	13632,2	3208,3	0,002	39
22	2	1r1o	0,003	0,001	0,008	0,004	0,92	0,10	0,00	-	1,11	0,32	1,12	0,43	0,21	-	33,1	13,0	12801,8	3376,1	13623,7	2913,7	13544,2	2972,7	0,003	32
19	2	1r1o	0,003	0,001	0,008	0,003	0,94	0,12	0,01	-	2,04	4,14	2,06	3,90	0,43	-	41,5	18,7	11427,3	3963,5	13526,5	2814,7	13778,4	3347,1	0,044	10
16	2	1r1o	0,003	0,001	0,008	0,003	0,93	0,18	0,01	-	2,56	4,68	2,55	4,67	0,49	-	45,5	15,0	10505,3	4544,8	13390,7	3105,3	13417,4	3157,2	0,078	4
13	2	1r1o	0,002	0,001	0,008	0,003	0,95	0,17	0,08	-	5,13	7,98	5,28	7,91	0,76	-	55,9	11,9	7689,7	4903,8	13693,6	3387,7	14291,8	3707,7	0,102	1
33	2	2o	0,003	0,001	0,008	0,003	0,70	0,35	0,00	-	1,30	0,46	1,84	1,26	0,56	-	81,6	93,7	7706,2	2681,8	9264,1	2360,2	11882,9	3889,6	0,006	26
32	2	2o	0,003	0,002	0,007	0,002	0,75	0,32	0,03	-	1,72	3,00	2,36	4,63	0,59	-	67,7	61,5	7746,0	2840,7	9564,1	2642,1	12057,2	3882,1	0,013	21
30	2	2o	0,002	0,001	0,007	0,002	0,87	0,24	0,04	-	2,60	3,15	3,35	3,68	0,75	-	65,0	44,0	6798,6	3560,2	10491,1	2873,5	13185,0	3933,0	0,051	7
12	2	2o	0,002	0,002	0,008	0,002	0,80	0,31	0,08	-	3,44	5,14	4,00	5,20	0,72	-	87,1	84,1	6345,1	3471,3	10033,1	2996,1	12496,7	4144,0	0,047	8
31	2	2o	0,003	0,002	0,008	0,003	0,68	0,33	0,03	-	1,32	0,51	1,79	1,23	0,60	-	68,9	77,5	7305,1	2477,8	8971,4	2489,5	11250,4	3897,0	0,009	23
29	2	2o	0,003	0,002	0,008	0,003	0,70	0,35	0,03	-	1,97	2,24	2,64	3,04	0,70	-	63,5	61,5	6555,0	2838,5	9035,8	2356,3	11365,9	3548,1	0,042	11
11	2	2o	0,002	0,001	0,007	0,002	0,72	0,31	0,03	-	2,66	6,17	3,39	6,58	0,72	-	91,1	89,5	6538,5	2800,7	8898,6	2455,8	11836,7	3813,3	0,013	20
28	2	2o	0,003	0,002	0,008	0,003	0,67	0,32	0,04	-	1,44	0,76	2,46	2,02	0,67	-	83,0	86,0	5892,9	2344,1	7629,9	2411,5	11360,5	4401,1	0,008	24
10	2	2o	0,003	0,002	0,007	0,002	0,68	0,32	0,04	-	1,78	1,49	2,70	2,42	0,78	-	74,4	54,2	5871,8	2670,8	8118,4	2184,3	11099,4	3419,6	0,010	22
9	2	2o	0,003	0,002	0,008	0,003	0,66	0,31	0,06	-	1,50	1,39	2,50	2,43	0,74	-	81,2	84,7	5905,9	2235,4	7521,2	2188,6	11298,5	4283,9	0,004	29
35	4	3r1o	0,003	0,001	0,009	0,004	0,95	0,14	0,00	-	1,20	1,78	1,24	1,93	0,19	-	182,6	93,7	13895,0	4168,8	14290,4	3836,4	14186,3	3486,0	0,002	37
40	4	2r2o	0,003	0,001	0,008	0,004	0,95	0,15	0,00	-	1,06	0,19	1,11	0,48	0,22	-	138,3	61,0	12988,2	3472,6	13456,9	3177,2	13694,0	3376,0	0,000	42
37	4	2r2o	0,003	0,002	0,008	0,004	0,95	0,12	0,00	-	1,46	2,32	1,39	1,72	0,28	-	124,5	100,6	12695,8	4166,2	13645,3	3399,5	13616,6	3327,7	0,018	14
36	4	2r2o	0,003	0,002	0,008	0,004	0,93	0,18	0,00	-	1,33	1,25	1,45	1,33	0,39	-	82,2	55,9	11246,5	3757,8	12535,3	2919,4	13329,9	2928,6	0,018	15
39	4	1r3o	0,003	0,001	0,008	0,002	0,87	0,26	0,03	-	1,51	1,67	2,00	2,31	0,56	-	110,4	89,0	9390,5	3914,1	10922,2	3327,8	13139,0	3769,2	0,017	16
38	4	1r3o	0,003	0,001	0,008	0,004	0,91	0,21	0,02	-	2,38	4,23	2,43	4,05	0,50	-	56,2	16,5	10290,3	4331,4	12882,2	3299,4	13299,3	3516,5	0,074	6
34	4	4o	0,002	0,001	0,007	0,002	0,80	0,31	0,02	-	2,42	3,22	3,12	3,58	0,66	-	83,3	56,4	6733,3	3133,2	9925,5	2682,8	12767,2	4031,5	0,045	9
41	8	3r5o	0,003	0,001	0,008	0,003	0,89	0,24	0,02	-	1,40	1,59	1,60	1,80	0,38	-	93,7	68,2	11001,7	3965,3	12358,1	3369,6	13276,9	3254,4	0,013	19
42	16	6r10o	0,003	0,001	0,008	0,004	0,89	0,23	0,01	-	1,18	0,76	1,37	1,30	0,26	-	134,3	77,3	11372,6	3713,2	12184,8	3235,9	12829,8	3094,8	0,005	28

Tabella2.10: Stime dei parametri demografici e indici di qualità delle stime calcolati per ogni schema di campionamento nello scenario demografico N_0 1500 I_C 100. Vedi materiali e metodi per la descrizione degli indici. Sd: deviazione standard. Rank: classifica degli schemi di campionamento secondo P_{SS}.

CODICE	GRUPPO	SOTTOGRUPPO	RRE		RCI95%		COV95%		P_C		I1		I2		P_{RID}		T_C		N_0		N_{PRE}		N_{MRCA}		P_{SS}	Rank
			Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd	Media	sd		
8	1	controllo	0,004	0,002	0,005	0,003	0,97	0,09	0,00	-	1,00	0,00	0,96	0,13	0,02	-	272,5	232,6	150181,5	22526,4	150262,5	22449,7	143923,4	26085,4	0,001	35
7	1	controllo	0,004	0,002	0,005	0,003	0,96	0,10	0,00	-	1,01	0,04	0,97	0,14	0,05	-	121,2	169,1	152239,4	21330,3	152868,8	20487,4	147023,3	23572,1	0,003	29
6	1	controllo	0,004	0,002	0,005	0,006	0,96	0,07	0,01	-	1,00	0,01	1,00	0,20	0,06	-	3037,0	5418,3	150442,9	25148,7	150623,8	25050,3	147853,3	22165,5	0,001	36
5	1	1o	0,000	0,000	0,002	0,001	0,76	0,20	1,00	-	2,03	2,42	18,98	8,77	1,00	-	281,9	106,1	8714,7	3075,3	14652,1	9450,5	144464,0	30582,9	0,000	40
4	1	1o	0,000	0,000	0,002	0,001	0,79	0,16	1,00	-	2,09	2,21	25,70	10,98	1,00	-	263,3	104,5	6236,8	2443,8	11179,0	8072,9	139867,3	29291,6	0,003	30
3	1	1o	0,000	0,000	0,002	0,001	0,75	0,19	1,00	-	1,94	1,97	29,42	12,36	1,00	-	275,3	128,7	5326,5	1823,0	9292,4	8509,9	141929,6	36475,7	0,005	28
2	1	1o	0,000	0,000	0,002	0,001	0,72	0,19	1,00	-	2,17	3,25	31,26	12,45	1,00	-	267,4	101,5	4693,3	1492,8	8684,6	9029,7	134091,0	34379,2	0,000	38
1	1	1o	0,000	0,000	0,002	0,001	0,76	0,18	1,00	-	2,29	4,02	38,51	16,08	1,00	-	262,4	96,5	3940,1	1213,7	7406,6	7904,8	137582,6	33251,7	0,000	41
27	2	1r1o	0,003	0,002	0,004	0,002	0,96	0,14	0,01	-	1,17	0,55	2,63	4,38	0,46	-	215,8	199,6	112219,4	46823,6	116678,0	42128,1	147317,1	21455,6	0,030	16
24	2	1r1o	0,001	0,001	0,003	0,002	0,91	0,16	0,24	-	1,43	0,71	11,89	10,57	0,84	-	180,1	41,8	43130,9	48671,8	50317,1	48722,4	144364,2	24386,8	0,002	31
21	2	1r1o	0,001	0,001	0,003	0,002	0,91	0,13	0,56	-	1,32	0,30	19,24	13,35	0,95	-	181,8	40,1	25033,4	36950,0	29738,9	39339,5	146763,8	24894,7	0,001	34
18	2	1r1o	0,000	0,000	0,002	0,001	0,90	0,13	0,83	-	1,31	0,47	29,65	14,51	1,00	-	186,8	19,9	7060,8	8806,5	9395,7	11596,7	145837,4	29715,1	0,001	33
15	2	1r1o	0,000	0,000	0,002	0,001	0,87	0,16	0,96	-	1,23	0,30	35,89	14,50	1,00	-	197,9	25,4	4827,6	3228,4	5843,0	3745,6	146097,8	22760,8	0,000	42
26	2	1r1o	0,003	0,002	0,005	0,003	0,94	0,18	0,02	-	3,65	11,35	3,68	10,69	0,32	-	97,2	79,9	113010,4	42200,6	130979,9	19648,2	134434,7	23487,5	0,046	9
23	2	1r1o	0,002	0,002	0,004	0,002	0,93	0,20	0,16	-	14,45	22,99	14,99	22,50	0,68	-	101,9	147,3	67970,3	57729,2	129333,0	29805,9	145686,7	23258,3	0,051	6
20	2	1r1o	0,001	0,001	0,003	0,002	0,94	0,17	0,59	-	42,54	33,27	43,10	31,92	0,92	-	84,6	32,4	24868,4	43512,2	134203,0	29132,3	143487,8	20502,0	0,064	5
17	2	1r1o	0,000	0,001	0,003	0,001	0,95	0,13	0,76	-	52,39	28,94	54,68	29,28	0,95	-	91,3	10,1	10743,1	27433,3	137259,3	26265,1	148030,4	22835,0	0,072	3
14	2	1r1o	0,000	0,000	0,002	0,001	0,94	0,12	0,92	-	57,55	27,50	68,19	29,02	1,00	-	98,2	10,0	2493,8	1356,7	120248,5	27339,4	143246,3	24201,4	0,075	2
25	2	1r1o	0,003	0,002	0,004	0,002	0,92	0,19	0,04	-	1,05	0,18	1,23	0,70	0,21	-	395,6	900,0	129376,6	30341,3	132636,1	26526,0	142427,0	22137,1	0,000	39
22	2	1r1o	0,003	0,001	0,004	0,002	0,93	0,16	0,07	-	13,55	43,84	14,10	44,58	0,53	-	235,4	1083,4	98990,5	49740,8	126518,8	26277,1	145192,9	21364,5	0,008	25
19	2	1r1o	0,002	0,001	0,004	0,002	0,93	0,17	0,34	-	65,40	73,90	65,53	72,61	0,82	-	40,7	67,9	46522,4	55178,8	131011,4	29277,7	142266,2	21598,4	0,047	8
16	2	1r1o	0,002	0,001	0,004	0,002	0,96	0,11	0,53	-	90,27	76,84	85,66	70,58	0,92	-	58,7	176,7	23928,5	41738,3	137575,8	26505,2	138371,1	19713,1	0,085	1
13	2	1r1o	0,001	0,001	0,003	0,001	0,96	0,14	0,77	-	104,08	58,33	104,25	56,30	0,99	-	49,5	8,5	6743,9	19523,9	139838,0	23486,6	144047,9	23361,4	0,027	17
33	2	2o	0,000	0,000	0,002	0,001	0,78	0,18	1,00	-	4,26	8,97	21,31	11,96	1,00	-	292,0	143,5	8591,5	3861,6	20891,8	25883,8	148514,1	30192,4	0,007	27
32	2	2o	0,000	0,000	0,002	0,001	0,76	0,20	0,99	-	17,27	46,31	33,80	43,02	1,00	-	259,9	145,1	7246,7	4148,0	31110,9	40193,3	140662,9	29073,6	0,011	23
30	2	2o	0,000	0,000	0,002	0,001	0,82	0,19	1,00	-	36,14	64,97	55,65	60,23	1,00	-	187,7	106,1	4820,7	3067,0	50916,5	54607,9	144454,1	31098,0	0,041	12
12	2	2o	0,000	0,000	0,002	0,001	0,81	0,20	1,00	-	57,49	66,41	76,30	63,32	1,00	-	149,2	101,5	3534,1	2988,0	71467,9	56648,2	141444,8	27162,4	0,049	7
31	2	2o	0,000	0,000	0,002	0,001	0,75	0,18	1,00	-	3,87	11,25	26,45	12,60	1,00	-	275,4	124,7	6418,9	2955,1	14826,4	25613,7	143373,3	27489,6	0,001	32
29	2	2o	0,000	0,000	0,002	0,001	0,74	0,18	1,00	-	14,37	35,55	38,79	37,32	1,00	-	252,6	165,1	5214,8	2933,3	26195,4	35543,3	137956,6	32323,4	0,010	24
11	2	2o	0,000	0,000	0,002	0,001	0,77	0,17	1,00	-	37,67	77,63	64,48	73,13	1,00	-	195,7	113,2	3665,4	2345,6	43797,6	47964,2	138956,8	30307,0	0,066	4
28	2	2o	0,000	0,000	0,002	0,001	0,72	0,20	1,00	-	7,34	23,97	33,27	25,54	1,00	-	265,3	116,7	5071,5	2016,0	15895,8	28711,9	134737,6	28923,4	0,007	26
10	2	2o	0,000	0,000	0,002	0,001	0,76	0,21	1,00	-	19,18	40,30	49,57	39,76	1,00	-	245,6	159,1	4309,4	2561,4	29149,3	38877,9	145824,3	33716,5	0,012	21
9	2	2o	0,000	0,000	0,002	0,001	0,74	0,20	1,00	-	5,91	13,70	39,34	20,29	1,00	-	289,4	168,8	4479,0	2082,8	13747,1	20397,8	142716,0	34526,6	0,001	37
35	4	3r1o	0,003	0,002	0,005	0,003	0,96	0,14	0,01	-	10,56	38,19	8,76	31,03	0,30	-	151,5	308,3	119047,8	46735,5	137223,3	22521,5	140315,9	25545,0	0,031	15
40	4	2r2o	0,002	0,002	0,004	0,002	0,94	0,18	0,05	-	6,25	14,77	7,30	15,79	0,54	-	659,9	2523,3	96698,4	53661,8	124012,5	36182,2	149790,2	32081,9	0,012	22
37	4	2r2o	0,002	0,002	0,004	0,003	0,91	0,23	0,06	-	12,89	43,91	14,00	44,05	0,54	-	145,9	177,4	91870,3	52513,9	121237,6	32382,8	144118,8	25685,0	0,013	20
36	4	2r2o	0,001	0,001	0,003	0,001	0,92	0,18	0,48	-	62,64	61,86	63,91	59,26	0,87	-	80,2	93,1	22758,3	38570,7	117653,0	39315,2	139975,9	30186,9	0,045	10
39	4	1r3o	0,000	0,000	0,002	0,001	0,87	0,15	0,85	-	29,53	56,42	46,47	54,06	1,00	-	173,2	122,4	7861,9	10296,2	47906,5	48608,6	141396,1	27395,0	0,036	14
38	4	1r3o	0,001	0,001	0,002	0,001	0,91	0,20	0,68	-	102,60	78,51	100,38	72,55	0,99	-	73,8	92,4	8256,6	18556,1	125859,4	35297,8	138498,8	24470,9	0,044	11
34	4	4o	0,000	0,000	0,002	0,001	0,82	0,20	0,99	-	53,93	66,59	77,07	70,40	1,00	-	189,4	157,4	4662,5	4637,3	62588,8	54258,1	142909,5	25837,3	0,027	18
41	8	3r5o	0,001	0,001	0,003	0,002	0,91	0,20	0,52	-	58,00	69,88	61,80	68,02	0,97	-	107,5	169,8	22174,0	36947,2	111419,8	44083,4	142988,3	23798,5	0,041	13
42	16	6r10o	0,001	0,002	0,003	0,002	0,94	0,14	0,32	-	33,75	63,72	36,29	62,10	0,84	-	110,4	101,6	45595,4	51219,1	113244,1	39283,5	142591,4	21641,4	0,021	19

APPLICAZIONE NUMERO TRE

ARTICOLO DI REVIEW SULL'ABC (*APPROXIMATE BAYESIAN COMPUTATION*), UNA METODOLOGIA STATISTICA PER L'INFERENZA DEMOGRAFICA IN MODELLI COMPLESSI

Titolo originale: "ABC as a flexible framework to estimate demography over space and time: some cons, mani pros"

Autori: Giorgio Bertorelle, Andrea Benazzo e Stefano Mona

Rivista scientifica: Molecular ecology, 19(13):2609-25

Anno di pubblicazione: 2010

In questa sezione della tesi, è riportato un riassunto dell'articolo scientifico descritto nel titolo e inserito nella sezione "ALLEGATO".

Introduzione

La genetica di popolazione si occupa di capire ed analizzare la variabilità genetica all'interno e tra le popolazioni. Agli esordi di questa disciplina scientifica, pochissimi dati empirici erano disponibili per essere analizzati, perciò i primi genetisti di popolazione si occupavano di sviluppare predizioni teoriche da semplici modelli popolazionistici per poi compararle con le scarse informazioni genetiche presenti all'epoca. In seguito, con l'avvento e lo sviluppo delle tecniche di PCR, la produzione di marcatori genetici non fu più un problema, e le analisi descrittive di variabilità genetica come PCA, AMOVA e autocorrelazione spaziale furono ampiamente usate per descrivere i pattern di variabilità genetica osservati e iniziare a confrontare ipotesi evolutive. Più recentemente, l'incremento prestazionale dei personal computer e lo sviluppo della teoria Coalescente, hanno premesso lo sviluppo di una metodologia statistica chiamata ABC (Approximated Bayesian Computation, Beaumont et al. 2002) con la quale, per la prima volta, si sono potuti analizzare modelli demografici complessi (e quindi realistici) utilizzando la grande abbondanza di dati genetici disponibili. L'idea alla base di questa tecnica, prevede la simulazione di milioni di genealogie secondo diversi modelli di evoluzione e diversi parametri associati al modello, in seguito le simulazioni che hanno prodotto dati di variabilità genetica (riassunti da statistiche descrittive, SuSt da qui in poi) molto vicine ai dati osservati vengono conservate e studiate. I parametri demografici associati a queste simulazioni successivamente andranno a formare, dopo opportuni aggiustamenti, le distribuzioni a posteriori che verranno studiate nel dettaglio. La grande flessibilità di questa tecnica, nella sua versione bayesiana, permette di stimare le distribuzioni a posteriori dei parametri di interesse, il confronto tra modelli di evoluzione e la stima quantitativa della qualità dei risultati. Purtroppo, anche se alcuni pacchetti software sono disponibili, l'ABC non è semplice da essere utilizzata. L'utente è chiamato ad applicare con attenzione ogni fase prevista dall'ABC dato che un consenso generale su come procedere non è stato ancora raggiunto e, inoltre, deve eseguire un controllo della qualità dei risultati finali ottenuti. In conclusione, l'ABC è un'analisi dispendiosa dal punto di vista computazionale e piuttosto complicata da applicare, ma molto flessibile e potente. In questa review sono state descritte e commentate le fasi previste dall'ABC, sono stati analizzati alcuni delle applicazioni principali dell'ABC e sono stati forniti consigli per l'analisi ai potenziali utilizzatori.

Origini storiche e la definizione formale di ABC

La prima applicazione dell'idea alla base dell'ABC si trova in due lavori pubblicati nel 1997 (Fu e Li 1997; Tavare et al. 1997), dove gli autori stimarono la distribuzione a posteriori dell'età dell'antenato comune più recente (TMRCA) degli individui di una popolazione, partendo da dati genetici. Gli autori proposero di eseguire numerose simulazioni dal modello demografico considerato e di stimare la distribuzione del TMRCA da quelle con un indice di variabilità genetica uguale a quello osservato nel campione. Questi due lavori possono essere considerati i capostipiti di questa tecnica perché Fu e Li (1997) suggerirono un metodo in teoria estendibile a qualsiasi modello demografico, mentre Tavare et al. (1997) ebbero il merito di introdurre la componente bayesiana nell'analisi, un aspetto fondamentale nell'ABC moderno. In seguito, Weiss e von Haeseler (1998) proposero di usare più di una SuSt per la stima dei parametri demografici e l'idea che anche le simulazioni che non producevano valori di variabilità genetica uguale (ma molto vicina) potevano essere utilizzate nell'analisi. Inoltre, per la prima volta in questo lavoro, l'ABC venne utilizzato anche per confrontare diversi modelli demografici, idea che poi fu meglio formalizzata ed estesa da Pritchard et al. (1999) con la piena integrazione della componente bayesiana nell'ABC. Da questo momento, questa tecnica si diffuse molto velocemente grazie al fatto che la stima dei parametri demografici di un modello poteva essere effettuata senza dover conoscere la funzione di likelihood del modello, ma semplicemente approssimandola in modo empirico attraverso l'uso delle simulazioni. Infatti, quando i dati genetici sono sostituiti da SuSt, la distribuzione a posteriori ricostruita è $P(\theta|\rho(SuSt_{SIM},SuSt)\leq\varepsilon)$, dove ρ è una qualsiasi misura di distanza tra le SuSt simulate ed osservate e ε è una soglia arbitraria. Per ε->0 (e se SuSt è sufficiente) la distribuzione approssimata coincide con $P(\theta|D)$, perciò per piccoli valori di ε si raggiunge un buon rapporto tra accuratezza ed efficienza. Beaumont et al. (2002) ebbero il merito di formalizzare e generalizzare l'ABC. Il miglioramento principale che proposero, fu quello di usare una modello di regressione lineare pesata per descrivere la relazione tra i parametri associati alle simulazioni con minore distanza dai dati osservati e le SuSt calcolate. L'idea fu quella di modificare il valore dei parametri in modo da riprodurre il caso in cui tutte le SuSt associate ai parametri fossero uguali a quelle osservate nei dati reali. Questa fase, introdusse una miglioria sia nella stima dei parametri, sia nell'efficienza del metodo in generale, di fatto riducendo il numero di simulazioni da eseguire potendo usare valori di ε maggiori. In seguito sono state proposte modifiche di singole parti della metodologia (ad esempio l'approccio GLM di Leuenberger e Wegmann 2010) o modifiche più radicali dell'algoritmo di approssimazione come l'approccio MCMC (Marjoram et al. 2003) o "Sequential Monte Carlo" (Doucet et al. 2001).

ABC in nove fasi

In questa review è stato esteso e generalizzato lo schema delle diverse fasi necessarie per compiere un'analisi ABC completa. Lo schema (Figura3.1) si compone di nove parti principali, per le quali segue una breve descrizione.

Fase1: il modello demografico

La storia e la demografia delle popolazioni da analizzare, e dai loro parametri associati, devono essere specificati insieme ai parametri genetici dei loci analizzati. In questa fase si possono definire modelli di evoluzione delle popolazioni complicati a piacimento, inserendo anche popolazioni non campionate ed eventi demografici come riduzioni, espansioni, estinzioni o traslocazioni. L'unico limite è la capacità dei simulatori di riprodurre il modello.

Fase2: l'incorporazione dell'informazione a priori

L'ABC è una metodologia di analisi bayesiana, perciò è possibile includere al suo interno le informazioni a priori disponibili sui modelli che devono essere studiati e sui loro parametri demografici. Queste informazioni sono utilizzate insieme a quelle contenute nei dati genetici per ottenere le distribuzioni a posteriori. Nella fase di simulazione, le distribuzioni a priori saranno disegnate (forma ed ampiezza della distribuzione) in modo da integrare quanto di noto su ogni particolare parametro e anche la probabilità a priori di ogni modello viene assegnata in rapporto al numero di simulazioni fatte per ogni modello.

Fase3: scelta delle statistiche descrittive (SuSt)

La procedura dell'ABC nella sua totalità si basa sul confronto tra i dati genetici simulati e osservati. Questo confronto non è eseguito sui dati completi, ma su una serie di statistiche descrittive (SuSt) che ne riassumono l'informazione. Sfortunatamente, non si hanno informazioni su quante e quali statistiche debbano essere utilizzate in un'analisi. Quanto formalizzato nella descrizione dell'ABC è che le statistiche devono essere *sufficienti*, cioè portare a una distribuzione a posteriori identica a quella ottenibile usando i dati genetici completi. Questo principio è strettamente dipendente dal modello analizzato, dai suoi parametri e dai dati scelti, per cui alcune analisi sulla scelta delle statistiche devono, in generale, essere condotte.

Fase4: la simulazione del modello/i

Milioni di simulazioni dovrebbero essere prodotte per ognuno dei modelli definiti nella *Fase1*, ognuna delle quali ottenuta con una diversa combinazione di parametri estratta dalle

distribuzioni a priori scelte. Questa fase è quella che richiede il tempo di calcolo maggiore. I dataset generati di solito vengono archiviati in database di referenza, salvando solamente i valori dei parametri usati nella simulazione e le SuSt associate, calcolate sui dati completi. Il vantaggio dell'approccio standard dell'ABC è la possibilità di utilizzare questi database per la fase di inferenza dei dati osservati ma anche per ottenere dei dati pseudo-osservati, detti anche *pods*, che possono essere utilizzati per calcolare degli indici di qualità dell'analisi. In principio, simulatori in backward o in forward possono essere utilizzati in questa fase, ma a causa del numero di simulazioni da produrre, solo la prima tipologia sembra aver raggiunto l'efficienza richiesta in termini di velocità.

Fase5: filtro delle simulazioni

Una simulazione viene "selezionata" quando la distanza (scelta a piacere) multivariata tra le SuSt osservate e simulate è al di sotto di un certo valore soglia. In genere, la distanza Euclidea viene comunemente usata come misura di distanza e solo una piccola parte (in genere dall'1 al 3% del totale) delle simulazioni più vicine all'osservato viene selezionata. Il valore della soglia è scelto in maniera arbitraria e deve perciò essere validato per verificare la robustezza delle stime al variare della soglia o utilizzando i *pods*.

Fase6: scelta del modello

L'ABC permette di confrontare ipotesi differenti sull'evoluzione di un processo attraverso l'assegnazione, ad ognuna di queste, di una probabilità a posteriori. Anche in questo caso, informazioni a priori in possesso dell'utente possono essere introdotte nell'analisi, in particolare, variando il numero di simulazioni prodotte secondo ognuno dei modelli, dove un maggior numero di simulaizoi corrisponde ad un maggior supporto a priori. La probabilità a posteriori di un modello è informativa di quanto i dati supportino ogni particolare modello, e sommando ad uno, permette il confronto tra loro. Questa probabilità può essere calcolata in diversi modi: stimata dalla frazione di simulazioni generate da ciascun modello nell'insieme delle simulazioni selezionate (Pritchard et al. 1999); eseguendo una regressione logistica multinomiale pesata (Beaumont 2008) sulle simulazioni selezionate tra tutti i modelli, e dove la variabile risposta è una variabile categorica indicatrice del modello, le variabili esplicative sono le SuSt delle simulazioni e i pesi sono attribuiti sulla base della distanza tra le SuSt simulate ed osservate; attraverso l'uso di un *General Linear Model* (Leuenberger e Wegmann 2010). Il Bayes Factor, un'altra misura per riassumere le evidenze in favore di un modello, è altrettanto facilmente calcolabile in un'analisi ABC.

Fase7: il controllo della qualità nella scelta del modello

L'ABC, nel suo insieme, può essere utilizzata per studiare la robustezza dell'intero apparato di scelta del modello. *Pods* simulati secondo precise combinazioni di parametri e modelli demografici possono essere utilizzati per valutare l'errore nella scelta del modello migliore. L'errore di I e II tipo può essere calcolato usando i vari modelli, di volta in volta, come ipotesi nulla o alternativa e registrando il supporto attribuito dall'ABC nei vari casi. Queste misure si possono rivelare estremamente utili per valutare quanto essere confidenti nel supportare i risultati ottenuti. Inoltre, è possibile utilizzare i *pods* per calcolare media e varianza delle probabilità a posteriori di ogni modello o ragionare sugli intervalli di credibilità dei modelli per avere un quadro generale sulle performance della fase di scelta del modello nella sua totalità

Fase8: stima dei parametri

L'ABC permette di ottenere le distribuzioni a posteriori dei parametri associati al modello oggetto di studio. Una volta selezionate le simulazioni che producono i dati genetici più vicini a quelli osservati, i parametri che hanno prodotto quelle simulazioni vengono comunemente modificati attraverso una regressione lineare pesata per la distanza tra le SuSt simulate e osservate. I parametri vengono modificati in modo da riprodurre il caso in cui tutte le tutte le SuSt coincidano con i valori osservati, ed è stato dimostrato che questa fase non solo aumenta la precisione della stima ma permette di usare valori di soglia più permissivi, che si traduce nel vantaggio di dover produrre meno simulazioni. Più recentemente è stato proposto da Leuenberger e Wegmann (2010) di utilizzare un General Linear Model per descrivere la relazione tra le SuSt selezionate e i parametri associati, ed ottenere le distribuzioni a posteriori per ogni parametro. Generalmente, la media, moda e mediana sono utilizzate come stimatori puntuali delle distribuzioni a posteriori, anche se non esiste accordo su quale sia lo stimatore con minor bias e varianza, e l"High posterior density interval" (analogo bayesiano dell'intervallo di confidenza) come misura di confidenza della stima.

Fase9: controllo di qualità del modello e della stima dei parametri

Oltre che per la scelta del modello, l'ABC permette di valutare la qualità delle stime ottenute. Un prima misura di qualità è la proporzione della varianza dei parametri spiegata dalle SuSt calcolate. Se questa misura, chiamata comunemente *coefficiente di determinazione*, è piccola, allora è difficile immaginare di poter stimare in maniera accurata le distribuzioni a posteriori dei parametri. Per valutare meglio la qualità delle stime, vengono impiegati nuovamente i *pods*. Le stime puntuali ottenute per ogni parametro sono utilizzate per generare un certo numero di

simulazioni, che saranno a loro volta utilizzate come dati osservati per la stima dei parametri. In questo modo, conoscendo i parametri reali che dovrebbero essere stimati, è possibile calcolare numerosi indici di qualità della stima. Inoltre, è possibile eseguire un *Posterior predictive test* (Gelman et al. 2003) per verificare quanto la combinazione di uno specifico modello e la stima dei parametri associati riesca a riprodurre i dati genetici osservati. Quest'ultima analisi può essere eseguita considernado le SuSt nel loro insieme o analizzando ogni singola SuSt in modo da identificare quali non siano riprodotte correttamente.

Applicazioni

In questa review sono riassunte le informazioni su 107 articoli pubblicati, dove l'ABC viene applicata dal 1997 al 2010, indicando quali sono stati gli organismi oggetto di studio, i marcatori genetici più usati, gli approcci più utilizzati per la scelta del modello e per la stima dei parametri, e infine, le 152 distribuzioni a priori e a posteriori dei parametri stimati in 14 lavori sono state utilizzate per verificare come attraverso l'ABC sia diminuita l'incertezza su parametri importanti riguardanti le specie oggetto di studio.

Conclusioni

L'Approximate Bayesian Computation ha il potenziale per diventare uno strumento standard per l'analisi di modelli demografici complessi applicati a dati genetici. In questa review sono stati descritti gli aspetti teorici e tecnici riguardanti questa metodologia insieme ai vantaggi e agli svantaggi che la sua applicazione comporta. Uno dei vantaggi principali è quello di permettere l'inferenza anche quando la likelihood del modello non è conosciuta ed è perciò possibile analizzare modelli realistici di evoluzione delle popolazioni intrattabili con metodi tradizionali basati sulla likelihood. Una volta eseguita l'analisi, è possibile inoltre eseguire una serie di controlli per valutare la qualità del processo inferenziale con un limitato dispendio aggiuntivo di risorse computazionali. Il fatto di dover controllare se l'analisi produce dei risultati credibili viene bilanciata dalla possibilità di riutilizzare le simulazioni salvate nel database di referenza, che in teoria potrebbero essere condivise tra gli utenti in modo da diminuire maggiormente il tempo complessivo per l'analisi dei dati genetici di una specie. Uno degli svantaggi di questa metodologia è che non esiste un singolo software, in grado di eseguire tutte le fasi richieste dall'ABC, che si adatti a qualsiasi caso specifico. Tuttavia, sono disponibili diversi strumenti che se combinati in una pipeline,

possono formare uno strumento adatto all'analisi di potenzialmente qualsiasi tipo di modello demografico e dato genetico. Non bisogna però dimenticare che l'ABC è un metodo approssimato per definizione, il cui livello avrebbe bisogno di essere studiato nel dettaglio rispetto a metodi basati sulla likelihood. Purtroppo, questo confronto sarebbe possibile soltanto per modelli dove questa quantità è derivabile, cioè escludendo la quasi totalità dei modelli definiti come "complessi". Fortunatamente, l'ABC è in grado di auto valutare la qualità della sua inferenza rendendo di fatto possibile l'analisi di quei modelli non analizzabili con altri metodi. Quando si decide di utilizzare l'ABC, oltre al tempo necessario alla produzione del database di referenza, una parte di tempo va investita nella definizione dei modelli e nell'ottimizzare ognuna delle fasi d'analisi, come ad esempio la scelta delle SuSt, la soglia da utilizzare, e il numero di simulazioni. Per questi motivi questa metodologia statistica richiede tempo e pazienza, ma la crescente capacità elaborativa e di archiviazione sta progressivamente riducendo il tempo necessario per la fase simulativa. Inoltre, alcuni metodi per ridurre il numero di simulazioni (ad esempio ABC con MCMC) sono in via di sviluppo, ma le loro performance devono essere ancora valutate nel dettaglio.

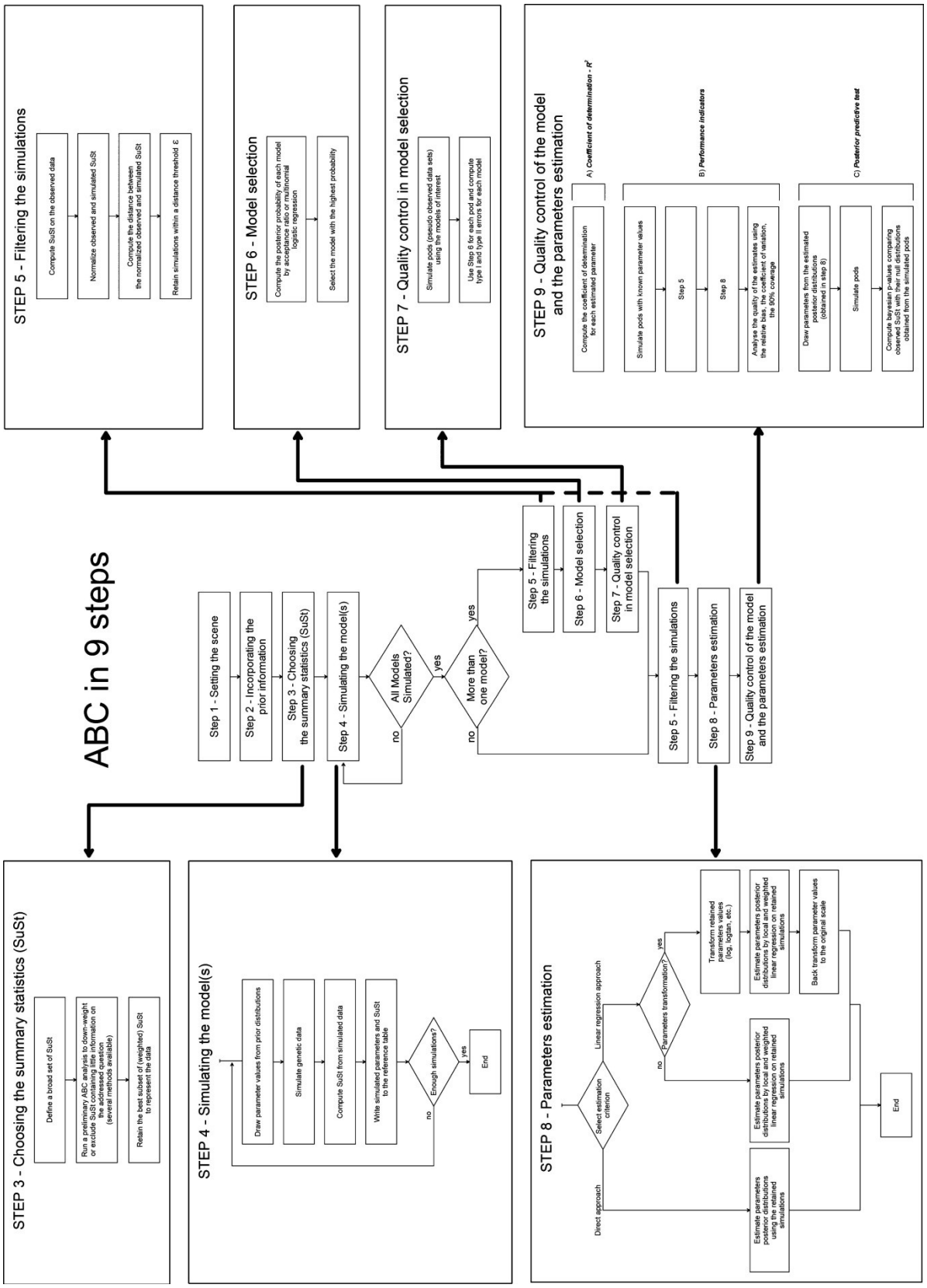

Figura 3.1: rappresentazione schematica delle fasi necessarie per un'analisi ABC.

APPLICAZIONE NUMERO QUATTRO

LA STORIA DEMOGRAFICA DEL GENERE *CHIONODRACO*, INFERITA ATTRAVERSO APPROXIMATED BAYESIAN COMPUTATION

4.1 INTRODUZIONE

L'Oceano del Sud ospita molti dei più estremi habitat presenti sulla Terra, e contiene al suo interno una fauna estremamente adattata al freddo, tra cui 129 specie di pesci perciformi appartenenti al sottordine dei *Notothenioidei* (Eastman 2005). La maggior parte dei *Notothenioidei* sono endemici delle acque antartiche, e hanno come areale di distribuzione le aree costiere ad alta latitudine, dove sono superiori in termini di abbondanza di specie e biomassa (Eastman e McCune 2000). Recentemente, è stata proposta un'ipotesi temporale, secondo la quale la notevole diversificazione ecologica e filogenetica che si osserva oggi nelle acque Antartiche si sia formata almeno 10 milioni di anni dopo l'origine dei *Notothenioidei* antartici, durante gli eventi glaciali nel tardo Miocene (11.6-5.3 milioni di anni fa). Questa teoria sottolinea quindi il ruolo fondamentale della diminuzione della temperatura, e dei cicli glaciali, nel creare le condizioni ecologiche necessarie alla speciazione (Near et al. 2011). All'interno del sottordine, la famiglia dei *Channichthyidae* è composta da 11 generi, la maggior parte dei quali monospecifici, con un totale di 16 specie conosciute. Una delle poche eccezioni è rappresentata dal genere *Chionodraco*, per il quale tre specie sono state descritte: *Chionodraco hamatus*, *Chionodraco myersi*, e *Chionodraco rastrospinosus* (Eastman e Eakin 2000). Le tre specie mostrano una morfologia estremamente simile, sebbene l'analisi del DNA mitocondriale indichi la presenza di tre gruppi monofiletici distinti (Patarnello et al. 2003). In accordo con quanto osservato per gli altri gruppi di specie antartiche, è stato ipotizzato che la periodica espansione dello strato di ghiaccio abbia promosso eventi di speciazione allopatrica, o di estinzione, come conseguenza della variazione nella disponibilità degli habitat (Janko et al. 2011), ma successivamente, il progressivo scioglimento dei ghiacci possa aver favorito eventi di ibridazione inter-specifica (Stelkens et al. 2010).

I metodi per l'inferenza demografica basati sulla likelihood, sono attualmente uno dei più sofisticati strumenti per studiare la storia demografica di una o più popolazioni. Partendo da un campione di DNA, è possibile, ad esempio, studiare se una popolazione ha avuto un singolo

aumento o riduzione della dimensione effettiva (MSVAR, Beaumont 1999) o ricostruire com'è variata la sua dimensione effettiva nel tempo (Skyline Plot, Drummond et al. 2005). Anche nel caso di popolazioni multiple, è possibile studiarne la storia demografica, analizzando nel dettaglio gli eventi di divergenza (IM, Hey 2010) o i tassi di migrazione (MIGRATE, Beerli e Felsenstein 2001). Ogni metodologia ha alla base un modello demografico e un modello mutazionale, che regolano rispettivamente l'evoluzione delle popolazioni nel tempo e l'evoluzione dei dati genetici. Una caratteristica fondamentale di queste metodologie è che la likelihood del modello (demografico + mutazionale) deve essere conosciuta, di fatto escludendone l'applicabilità se si vuole studiare un complesso modello (e quindi realistico) di evoluzione influenzato dai cicli glaciali nelle tre specie del genere *Chionodraco*. L'Approximate Bayesian Computation (ABC, Beaumont et al. 2002) è una metodologia statistica che permette di evitare le limitazioni imposte dal calcolo della likelihood e di ottenere precise stime di parametri demografici utilizzando realistici modelli di evoluzione (Fagundes et al. 2007; Neuenschwander et al. 2008). Inoltre, è stata precedentemente utilizzata con successo in numerosi studi per verificare in che modo eventi climatici come le glaciazioni abbiano modificato il pattern di variabilità genetica di specie animali (Neuenschwander et al. 2008; Row et al. 2011; Sjodin et al. 2012) e vegetali (Francois et al. 2008; Holliday et al. 2010).

Lo scopo di questo studio è di analizzare, utilizzando l'ABC, l'intensità di possibili eventi d'ibridazione, e la loro durata temporale, all'interno del genere *Chionodraco*. I campioni sono stati raccolti tra il 1989 e il 2007 dal professor Lorenzo Zane e tipizzati geneticamente nel suo gruppo di ricerca all'Università di Padova. In questo lavoro sono stati confrontati diversi scenari evolutivi che si differenziano per l'effetto dei due periodi interglaciali più recenti, l'Olocene e l'Eemiano e, i parametri demografici principali come le dimensioni effettive delle popolazioni e i tassi di migrazione sono stati stimati.

4.2 MATERIALI E METODI

I campioni e i marcatori genetici

I campioni dalle popolazioni di *C. hamatus*, *C. myersi* and *C. rastrospinosus*, sono stati raccolti tra il 1989 e il 2007 in quattro regioni geografiche antatiche: il Mare di Weddell, il Mare di Ross (Baia Terranova) e la Penisola Antartica (Isola Elephant e Isola Joinville, Tabella4.1). Gli esemplari sono stati in seguito assegnati a una delle tre specie sulla base dei caratteri mofologici seguendo le indicazioni fornita da Fisher & Hureau (1985). Otto microsatelliti sono stati isolati e tipizzati geneticamente in totale di 108 campioni (37 per *C. myersi*, 39 per *C. rastrospinosus* and 32 per *C. hamatus*).

Tabella4.2: Dimensioni e luogo di origine delle popolazioni campionate.

Specie	Luogo di campionamento	Numero di esemplari
C. hamatus	Mare di Weddell	9
	Mare di Ross (Baia Terranova)	23
C. myersi	Mare di Ross (Baia Terranova)	27
	Mare di Weddell	10
C. rastrospinosus	Isola Elephant	19
	Isola Joinville	20

Approximate Bayesian Computation (ABC)

L'approccio ABC standard, implementato nel software ABCtoolbox (Wegmann et al. 2010) ed integrato con script per l'ambiente statistico R (R Development Core Team 2010), è stato utilizzato per stimare la probabilità a posteriori di cinque differenti scenari evolutivi del genere *Chionodraco* e per ricostruirne poi le distribuzioni a posteriori dei parametri demografici. L'ABC sta diventando nel corso degli anni una metodologia standard per l'analisi di modelli complessi dove il calcolo della likelihood non è possibile. Nella sua versione standard, questo metodo prevede la generazione di un grande numero di simulazioni genetiche con le stesse caratteristiche del campione reale da analizzare (tipo di marcatore e numero di individui), da ogni modello demografico definito . In ogni simulazione, i parametri demografici sono estratti da opportune distribuzioni a priori e un certo numero di statistiche descrittive sono in seguito calcolate sui campioni simulati. Le simulazioni originate da tutti i modelli demografici che producono la minima differenza tra le statistiche descrittive calcolate sulle simulazioni e sui dati osservati vengono utilizzate per inferire le probabilità dei modelli e stimare i parametri demografici. Per la definizione

formale del metodo e per la descrizione approfondita di ogni sua parte, si veda il capitolo "Applicazione tre" di questa tesi.

I modelli demografici simulati

Le simulazioni coalescenti sono state prodotte utilizzando il modulo simulativo contenuto in ABCtoolbox costituito dal software Simcoal2 (referenza). La definizione di un modello demografico e un modello mutazionale è necessaria per la fase simulativa. Il primo viene utilizzato per descrivere come le popolazioni si evolvono nel tempo, specificando in ogni istante temporale quale siano le dimensioni delle popolazioni ed eventuali eventi che ne influenzino la demografia come riduzioni improvvise o eventi migratori. Il secondo invece descrive la frequenza, e le caratteristiche associate, delle mutazioni nelle popolazioni. Per descrivere l'evoluzione del genere *Chionodraco* sono stati definiti cinque scenari demografici, descritti graficamente in Figura4.1 con i parametri associati. Tutti i modelli condividono la sequenza di divergenza tra le specie (*C. myersy*,(*C. rastrospinosus*, *C. hamatus*)), ma ognuno è caratterizzato da una differente storia migratoria. La linea evolutiva di *C. myersi* proviene dal primo evento di divergenza, fissato a 2 milioni di anni fa, mentre, la divergenza tra *C. rastrospinosus* e *C. hamatus* è stata fissata a 1.8 milioni. Questi tempi, stimati da Near et al. (2011), sono stati fissati in modo da ridurre il numero complessivo di parametri liberi di ogni modello, e anche perchè i microsatelliti non sono altamente informativi per le ricostruzioni filogenetiche. Ogni modello è caratterizzato da cinque dimensioni effettive: una per ogni popolazione moderna (*C. myersi* (N_1),*C.rastrospinosus* (N_2),*e C. hamatus* (N_3)), una per la popolazione ancestrale di *C. rastrospinosus* e *C. hamatus* (N_{A1}), e una per la popolazione ancestrale di tutte e tre le specie (N_{A2}). Nel modello M1 si assume il completo isolamento tra le specie, ciòe non sia mai avvenuto nessun evento di ibridazione tra le specie. Al contrario, nel modello M2, è stata simulato un livello costante di migrazione di individui tra le specie in modo da riprodurre il caso in cui l'ibridazione sia avvenuta ma non sia stata influenzata da nessun evento climatico. Nei modelli M3 e M4, la migrazione è stata resa possibile solo in uno nei due periodi interglaciali più recenti: il periodo Eemiano, tra i 110 000 e i 130 000 anni fa, e il periodo Olocenico da 10 000 anni fa ad oggi (Ivy-Ochs et al. 2008). Il modello M5 è una semplice combinazione di M3 ed M4, e dove perciò individui migranti possono essere scambiati nelle due finestre temporali interglaciali. Gli ultimi tre modelli (M3, M4, M5) sono stati realizzati in modo da studiare l'impatto degli ultimi due cicli glaciali sulle dinamiche demografiche delle tre specie. Un tempo di generazione di 6 anni (La Mesa e Vacchi 2001; La Mesa e Ashford 2008) è stato utilizzato per scalare tutti i tempi all'interno dei modelli, come documentato per il genere *Chionodraco* (inserire referenza). Il Generalized Stepwise Mutation Model (GSM, Estoup et al. 2002) è stato

utilizzato come modello mutaizonale per generare la variabilità genetica agli 8 loci microsatelliti. Questo modello oltre a regolare la frequenza con cui un microsatellite muta nel tempo, introduce un parametro aggiuntivo *P* che specifica il numero di ripetizioni inserite o eliminate da ogni evento mutazionale. Recentemente, sono state sollevate alcune critiche sulla capacità dell'ABC di poter discriminare tra modelli annidati (Templeton 2009; Templeton 2010), cioè quando un modello è un caso specifico di un modello più generale, per cui, per sicurezza, è stato deciso di evitare questa situazione. Tutti e cinque i modelli non sono annidati grazie alle distribuzioni a priori assegnate a ogni modello. Ad esempio il modello M1 sarebbe uguale al modello M2 nel caso specifico in cui i tassi migrazione tra le specie di M2 fossero uguali a 0. Escludere il valore di 0 dalle distribuzioni a priori (ad es Log-uniforme $10^{-8} – 10^{-2}$) dei tassi di migrazione è stato dunque sufficiente per evitare di incorrere in modelli annidati.

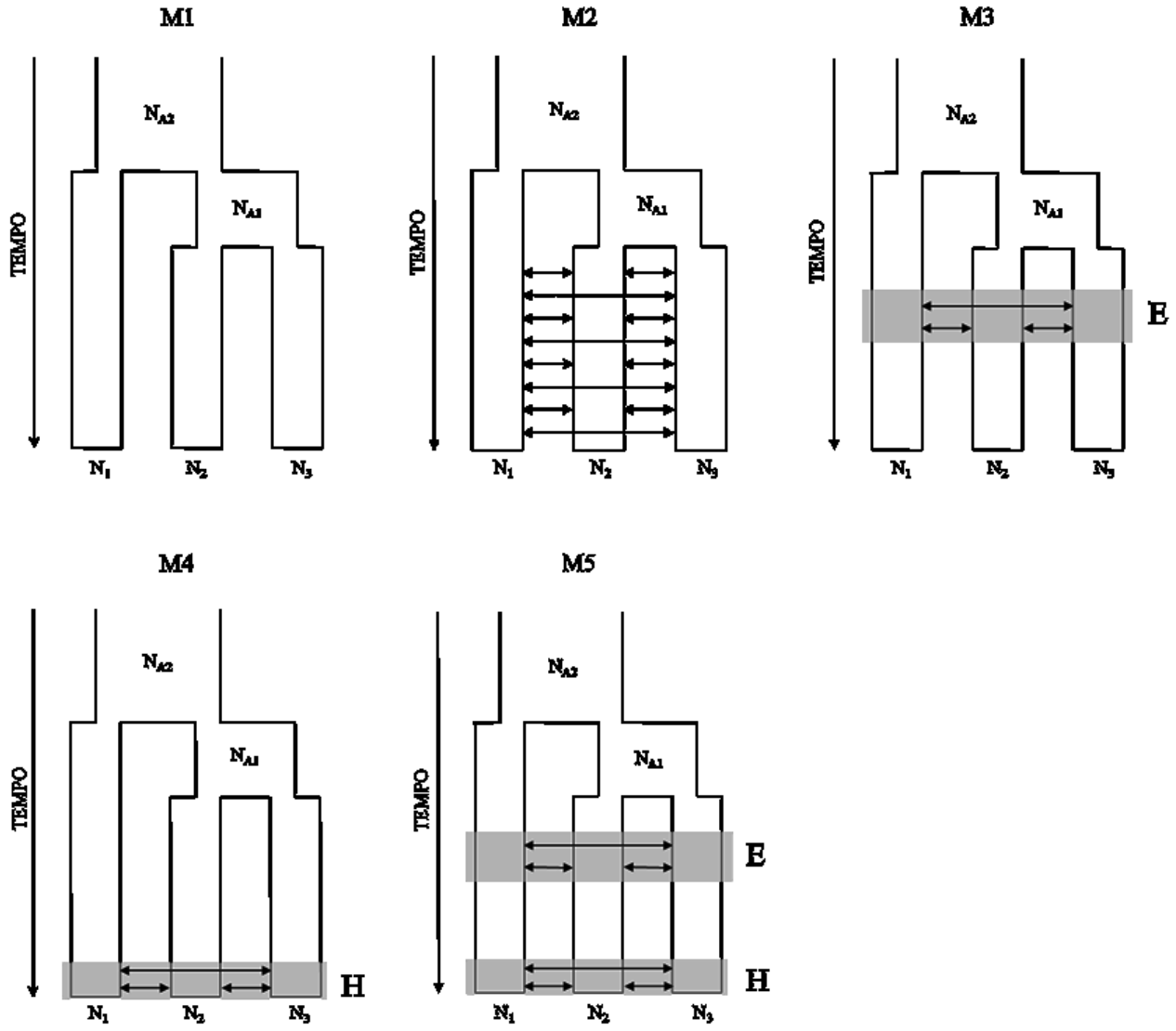

Figura4.1: Gli scenari demografici analizzati con le loro probabilità a posteriori stimate secondo i tre differenti approcci. Le frecce nere corrispondono ai flussi migratori tra le specie *C. myersi* (*Ne_M*), *C. rastrospinosus* (*Ne_R*) and *C. hamatus* (*Ne_H*). I periodi interglaciali Olocenico (H) ed Eemiano (E) sono evidenziati in grigio.

Distribuzioni a priori

Le distribuzioni a priori dei parametri di ogni modello sono state definite sulla base delle informazioni presenti in letteratura. A tutte le dimensioni effettive, moderne e ancestrali, è stata associata una distribuzione a priori uniforme nell'intervallo tra 250 e 20 000 individui diploidi (da 500 a 40 000 cromosomi), sulla base di precedenti stime effettuate in *Chaenocephalus aceratus* (Papetti et al. 2009), appartenente alla stessa famiglia del genere *Chionodraco*. Anche i tassi di migrazione condividono lo stesso tipo di distribuzione a priori. La probabilità che un individuo migri in un'altra popolazione è stata estratta da una distribuzione log-uniforme tra 10^{-8} a 10^{-2}. In questo modo, vengono favoriti bassi livelli di ibridazione, come plausibile quando si tratta di specie, ma allo stesso tempo sono possibili eventi di più grande impatto con minore probabilità. Per regolare il tasso di mutazione è stato seguito l'approccio a hyper-prior usato da Excoffier et al. (2005) e Neuenschwander et al. (2008). Per pima cosa è stata estratta la media del tasso di mutazione μ tra i loci da una distribuzione Normale, con media $5x10^{-4}$ e varianza $1.3x10^{-4}$, limitata nell'intervallo tra 10^{-7} e 10^{-3}. In seguito, i tassi di mutazione locus specifici sono stati estratti da una distribuzione Gamma con media μ e parametro di forma uguale a 2 (Excoffier et al. 2005; Neuenschwander et al. 2008; Guillemaud et al. 2010). Infine, lo stesso approccio è stato utilizzato per il parametro *P* che regola la distribuzione geometrica del numero di ripetizioni che un evento mutazionale può inserire o togliere ad un microsatellite. La media di *P* è stata estratta da una distribuzione uniforme nell'intervallo tra 0 e 0.8, mentre i valori locus specifici di *P* da una distribuzione Beta(a,b) dove $a = 0.5+199P$ e $b = (1-P)/P$ (Excoffier et al. 2005; Neuenschwander et al. 2008). Vedi Tabella4.2 per maggiori dettagli sulle distribuzioni a priori.

Tabella4.3: Distribuzioni a priori dei parametri demografici. N: dimensioni effettive (in numero di cromosomi) delle tre specie (1=*C. myersi*; 2=*C. rastrospinosus*; 3=*C.hamatus*) e delle popolazioni ancestrali (A1 e A2); m: tassi di migrazioni tra coppie di specie, dove E e H si riferiscono ai periodi interglaciali Eemiano e Olocenico; μ: tasso di mutazione medio; P: media della parametro che regola la distribuzione geometrica nel modello mutazionale; IR: irrilevante.

Parametro	Distribuzione	Media	Moda	Quantili				
				Min.	5,00%	50,00%	95,00%	Max.
N_1	Uniforme	19 750	IR	500	2 475	20 250	38 025	40 000
N_2	Uniforme	19 750	IR	500	2 475	20 250	38 025	40 000
N_3	Uniforme	19 750	IR	500	2 475	20 250	38 025	40 000
N_{A1}	Uniforme	19 750	IR	500	2 475	20 250	38 025	40 000
N_{A2}	Uniforme	19 750	IR	500	2 475	20 250	38 025	40 000
μ	Normale	$5x10^{-4}$	$5x10^{-4}$	10^{-7}	$3x10^{-4}$	$5x10^{-4}$	$7x10^{-4}$	10^{-3}
P	Uniforme	0.40	IR	0.00	0.04	0.40	0.76	0.80
m_{E12}	Log-uniforme	$7.2x10^{-4}$	10^{-8}	10^{-8}	$2.0x10^{-8}$	10^{-5}	$5.0x10^{-3}$	10^{-2}
m_{E13}	Log-uniforme	$7.2x10^{-4}$	10^{-8}	10^{-8}	$2.0x10^{-8}$	10^{-5}	$5.0x10^{-3}$	10^{-2}
m_{E23}	Log-uniforme	$7.2x10^{-4}$	10^{-8}	10^{-8}	$2.0x10^{-8}$	10^{-5}	$5.0x10^{-3}$	10^{-2}
m_{H12}	Log-uniforme	$7.2x10^{-4}$	10^{-8}	10^{-8}	$2.0x10^{-8}$	10^{-5}	$5.0x10^{-3}$	10^{-2}
m_{H13}	Log-uniforme	$7.2x10^{-4}$	10^{-8}	10^{-8}	$2.0x10^{-8}$	10^{-5}	$5.0x10^{-3}$	10^{-2}
m_{H23}	Log-uniforme	$7.2x10^{-4}$	10^{-8}	10^{-8}	$2.0x10^{-8}$	10^{-5}	$5.0x10^{-3}$	10^{-2}

Le statistiche descrittive utilizzate.

Nel framework dell'ABC, la likelihood è approssimata dalla distanza euclidea tra un insieme di statistiche calcolate sia sui dati simulati sia e su quelli reali. La scelta di queste statistiche è di fondamentale importanza per soddisfare la proprietà di sufficienza necessaria per raggiungere una buona approssimazione della likelihood. Sebbene alcuni metodi esistano per selezionare l'insieme di statistiche migliore (Joyce e Marjoram 2008; Nunes e Balding 2010), spesso sono troppo intensivi dal punto di vista computazionale e inoltre non assicurano che il set di statistiche sia "sufficiente". Un insieme di 29 statistiche descrittive indipendenti è stato calcolato sui dati osservati e simulati: la media e la deviazione standard dell'eterozigosità (Nei 1987) e del range allelico tra i loci, sono state utilizzate per descrivere il livello di variabilità interna alle specie, mentre la media di F_{ST} (Cockerham e Weir 1984), e un insieme di statistiche sul numero di alleli condivisi e non condivisi tra le specie, è stata utilizzata per riassumere la variabilità tra le specie. Il numero di alleli e la statistica M (Garza e Williamson 2001; Excoffier et al. 2005), due statistiche normalmente usate nell'analisi dei microsatelliti, sono state escluse dall'insieme delle statistiche descrittive in

quanto completamente correlate con le 29 statistiche definite in precedenza. In Tabella4.3 sono riportati i valori delle statistiche calcolate sul campione osservato con Arlequin3.5.1(Excoffier e Lischer 2010).

Le probabilità dei modelli e la stima dei parametri

Un milione di simulazioni sono state prodotte per ognuno dei cinque modelli demografici e la probabilità a posteriori di ognuno è stata calcolata mediante tre metodi differenti (vedi il capitolo "Applicazione tre" per maggiori dettagli sui primi due metodi). L'approccio diretto (REJ) (Pritchard et al. 1999), consiste nel calcolare la frazione di simulazioni appartenenti a ciascun modello all'interno dell'insieme delle 100 simulazioni che producono i dati più simili a quelli osservati. Il secondo approccio, introdotto da Beaumont (2008), consiste nell'utilizzare un modello di regressione logistica multinomiale pesata dove le variabili esplicative sono le simulazioni che hanno prodotto dati simili all'osservato e la variabile risposta è l'indicatore del modello di appartenenza di ogni simulazione. La regressione poi è pesata per la distanza di ogni simulazione dai dati osservati. Le migliori 50 000 simulazioni tra tutti i modelli sono state selezionate per questo calcolo. Il terzo metodo, proposto inizialmene da Fagundes et al. 2007ed esteso successivamente da Veeramah et al. 2011, prevede un aggiustamento delle probabilità a posteriori calcolate precedentemente con la regressione logistica multinomiale (R-ADJ). Dato un set di probabilità a posteriori osservate $x_i \ldots x_n$ (ad es quelle calcolate con il secondo approccio sui dati reali), è possibile calcolare la probabilità che un modello, M_A, sia il modello corretto tra tutti i $M_{i \ldots n}$ modelli:

$$\frac{\Pr(M_i = x_i \ldots M_n = x_n | \, PM_A = vero)}{\sum_i^n \Pr(M_i = x_i \ldots M_n = x_n | \, PM_i = vero)}$$

Per ognuno dei modelli, M_i, è stato stimato $\Pr(M_i = x_i \ldots M_n = x_n | \, PM_i = vero)$ generando 100 dataset pseudo osservati dal modello e stimando la densità multinomiale nel punto $x_i \ldots x_n$ utilizzando un kernel Gaussiano multidimensionale. Il bandwidth *h* necessario per il calcolo della densità è stato fissato ed uguale in tutte le dimensioni. Sebbene quest'approccio non sia ottimale (*h* dovrebbe essere stimato per ognuna delle dimensioni), è stato necessario a causa dell'eccessiva complessità del calcolo. Tuttavia, il bandwidth è stato variato nell'intervallo 0.05 – 0.3 per verificare la stabilità delle probabilità a posteriori al variare di *h*. Infine, per i primi due approcci ogni probabilità è stata ricalcolata variando il numero di simulazioni selezionate in modo da verificare l'influenza della soglia usata sulla stima della probabilità a posteriori. Per il metodo

diretto (REJ) sono state usate 100, 200 e 500 simulazioni mentre 20 000, 50 000 e 75 000 per il metodo basato sulla regressione logistica multinomiale.

Una volta identificato il modello migliore, la stima delle distribuzioni a posteriori dei suoi parametri demografici è stata eseguita selezionando le migliori 2 000 simulazioni (minore distanza dall'osservato) e applicando una regressione lineare locale pesata tra il parametri e le statistiche descrittive, associati alle simulazioni migliori. L'introduzione di questo aggiustamento, è stato dimostrato produrre migliori distribuzioni a posteriori (Beaumont et al. 2002). Purtroppo, questa fase può portare un parametro a uscire dall'intervallo della prior, per cui è stato deciso di applicare la trasformazione *logtan* ai parametri prima della fase di regressione (Hamilton et al. 2005) in modo da far rispettare ad ogni parametro l'intervallo della prior. I parametri sono stati successivamente ritrasformati nella scala originale dopo la regressione ed è stata calcolata la mediana e l'intervallo di credibilità (HPD) al 95% per caratterizzare ogni distribuzione a posteriori stimata.

Per la fase di scelta del modello e per la stima dei parametri, sono stati utilizzate le funzioni di R rese disponibli da Mark Beaumont (http://www.rubic.rdg.ac.uk/~mab/stuff/) e script appositamente sviluppati, eseguibili in R2.10 (R Development Core Team 2010).

Analisi di Potenza dell'ABC nella scelta del modello

Uno dei vantaggi dell'ABC consiste nella possibilità di verificare la robustezza dell'intero apparato inferenziale, investendo solamente una piccola quantità aggiuntiva di risorse di calcolo. Tramite l'utilizzo di pseudo dataset osservati (*pods*) è stato possibile calcolare la frazione di veri e falsi positivi associata ad ogni modello. 100 pods sono stati simulati, estraendo i valori dei parametri dalle distribuzioni a priori definite in precedenza, per ognuno dei modelli (M1-M5, 500 pods in totale). Ogni pods è stato poi analizzato come se fosse un dataset reale e la probabilità a posteriori dei modelli è stata stimata secondo l'approccio diretto o la regressione logistica multinomiale (REJ, REG). Il terzo approccio (R-ADJ), non è stato preso in considerazione in questa fase perché contiene già al suo interno un aggiustamento proporzionale alla probabilità di commettere un errore nella scelta del modello basato su *pods*. Per ogni modello, ad esempio M1, il tasso di veri positivi è stato stimato come la proporizone di pods, simulati secondo M1, che lo supportano correttamente, mentre il tasso di falsi negativi come la frazione di pods, simulati con uno degli altri modelli (M2-M5), che però favoriscono M1. Due soglie decisionali di probabilità, 0.7 e 0.5, sono state prese in considerazione per poter valutare quando un modello viene supportato dai dati. Una soglia di 0.5 significa, ad esempio, che un modello è il favorito quando la sua

probabilità a posteriori è uguale o maggiore a 0.5. Inoltre, questa analisi è stata estesa anche utilizzando una soglia relativa, dove il modello favorito è quello ad ottenere la probabilità a posteriori più alta rispetto agli altri.

4.3 RISULTATI

Variabilità genetica osservata del genere Chionodraco

L'analisi dei 108 individui campionati indica la presenza di polimorfismo all'interno delle specie. *C. rastrospinosus* e *C. hamatus* mostrano valori medi di eterozigosità simili tra loro e leggermente superiori a *C. myersy* (Tabella4.3, H), anche se queste differenze si annullano se si tiene in considerazione la deviazione standard tra i loci. Anche il range allelico sembra essere piuttosto omogeneo all'interno dei gruppi, con l'unica eccezione rappresentata da *C. rastrospinosus*, dove il range di variazione sembra essere spostato su valori lievemente superiori rispetto alle altre due specie (Tabella4.3, R, specie 2). Come atteso, in accordo con la topologia di divergenza fissata nei modelli demografici, le specie più simili geneticamente sono *C.rastrospinosus* e *C. hamatus*, per le quali si osserva la minore distanza genetica (F_{ST}2-3=0.10, Tabella4.3) e un numero elevato di alleli condivisi (Cond2-3=9). La coppia di specie C. *myersi* - *C. hamatus* è la più divergente, mostrando il più alto valore di distanza genetica e il minor numero di alleli condivisi (FST=0.21, Cond=5, Tabella4.3), mentre, contro le attese, *C. myersi* e *C. rastrospinosus* sono caratterizzate da una bassa distanza genetica e dal più alto numero di alleli condivisi (FST=0.12, Cond=13, Tabella4.3). Questa ultima osservazione non sembra essere in accordo con quanto atteso dalla sequenza di divergenze supportata dal DNA mitocondriale (*C. myersi* equidistante da *C. rastrospinosus* e *C. hamatus*), suggerendo una possibile ibridazione tra queste due specie.

Tabella4.4: statistiche di variabilità genetica intra e tra specie. I nomi delle specie sono abbreviati: 1=*C. myersi*, 2=*C. rastrospinosus* and 3=*C. hamatus.* H: eterozigosità; R: range allelico; Priv: allelli privati; Cond: alleli condivisi; F_{ST}: indice di differenziamento; ds: deviazione standard tra i loci.

H

Specie	Media	ds
1	0.61	0.26
2	0.71	0.21
3	0.71	0.15

R

Specie	Media	ds
1	12.20	10.10
2	16.20	10.60
3	13.31	9.66

Priv

Specie	Totale	ds
1	8	0.92
2	18	1.48
3	19	1.99

Cond

Specie	Totale	ds
1+2+3	52	5.37
1+2	13	2.00
1+3	5	0.91
2+3	9	0.61

F_{ST}

Specie	Media
1-2	0.12
1-3	0.21
2-3	0.10

Il modello demografico supportato dai dati

Il modello senza migrazione (M1) viene fortemente sfavorito dall'analisi ABC (Tabella4.4). Questo modello infatti ottiene una probabilità uguale, o prossima, a 0 a seconda del metodo utilizzato ed indipendente dalla soglia. Anche il modello dove la migrazione è permessa solo nel periodo interglaciale Eemiano, ottiene scarso supporto, con probabilità che al massimo raggiungono lo 0.03 e sempre inferiori al modello migliore. Il metodo diretto (REJ) favorisce in maniera netta il modello di migrazione continua nel tempo tra le specie (M2), mentre il metodo della regressione (REG) supporta un modello demografico di complessità maggiore (M5), dove la migrazione avviene esclusivamente nei periodi interglaciali Eemiano e Olocenico (Tabella4.4). Quest'ultima evidenza (cioè che la migrazione sia avvenuta in corrispondenza dei periodi interglaciali) sembra essere confermata anche dal più sofisticato dei metodi di stima. In questo caso, la probabilità del modello M2, che veniva preferito dal metodo diretto, si abbassa a solo il 7%, mentre la probabilità del modello M5 sale al 58% (Tabella4.4, metodo R-ADJ). Inoltre, la probabilità associata al modello dove la migrazione è avvenuta solo in tempi recenti nel periodo interglaciale Olocenico (M4) è la seconda più alta fra tutti i modelli (32%), suggerendo che una migrazione recente (Olocenica) abbia influenzato il pattern di variabilità genetica maggiormente rispetto alla migrazione più antica precedente (Eemiana). Le probabilità calcolate secondo l'approccio R-ADJ sono risultate essere particolarmente stabili al variare di *h* (Figura4.4), confermando da credibilità dei valori calcolati (vedi materiali e metodi). Il numero di simulazioni utilizzate per definire la soglia di ogni approccio non sembra influenzare i risultati, infatti, le probabilità a posteriori rimangono stabili in ogni metodologia utilizzata.

L'analisi del numero di veri e falsi positivi generati nella fase di scelta del modello, utilizzando pseudo dataset osservati generati secondo i diversi modelli, indica che la combinazione di statistiche descrittive e dell'approccio ABC impiegati hanno un potere sufficiente per identificare il modello corretto. Quando non si utilizzano soglie troppo elevate (ad esempio 0.7), il numero di veri positivi è sempre elevato mentre il frazione di falsi positivi è bassa o addirittura molto bassa (Tabella4.5). L'approccio della regressione logistica produce risultati più credibili in quasi tutti i casi analizzati rispetto al metodo diretto. Infatti, nel confronto tra i due approcci, la regressione logistica ottiene una frazione di veri postivi più elevata per tutte le combinazioni di modelli e soglie, ad esclusione del modello M1 (Tabella4.5) dove il metodo diretto sembra avere prestazioni leggermente superiori. Anche il numero di falsi positivi risulta essere lievemente minore rispetto all'altro metodo, confermando quanto atteso da studi comparativi precedenti (Beaumont 2008). Il modello M5, selezionato dall'ABC come modello migliore per descrivere l'evoluzione del genere

Chionodraco, è caratterizzato da una frazione di falsi positivi molto bassa: nei 400 dataset pseudo osservati simulati secondo i modelli M1, M2, M3 e M4 (100 dataset per ogni modello), il modello M5 è stato scelto erroneamente come modello migliore solo nel 10% dei casi (M5, FP=0.10, Tabella4.5).

Tabella4.5: probabilità a posteriori dei modelli calcolate secondo i tre approcci considerati. REJ: metodo diretto; REG: metodo della regressione logistica multinomiale; R-ADJ: metodo della regressione logistica multinomiale modificato secondo Veeramah et al 2011. I numeri tra parentesi indicano il numero di simulazioni impiegate nella stima della probabilità.

	MODELLI				
	M1	M2	M3	M4	M5
REJ (100)	0.000	**0.810**	0.020	0.020	0.150
REJ (200)	0.000	**0.820**	0.020	0.035	0.125
REJ (500)	0.000	**0.810**	0.018	0.052	0.120
REG (20 000)	4.26×10^{-12}	0.286	1.63×10^{-4}	0.217	**0.495**
REG (50 000)	6.67×10^{-10}	0.294	1.67×10^{-4}	0.257	**0.448**
REG (75 000)	1.92×10^{-9}	0.307	1.92×10^{-4}	0.257	**0.434**
R-ADJ (50 000)	0.000	0.070	0.030	0.320	**0.580**

Tabella4.6: analisi di potenza della capacità dell'ABC di discriminare tra i modelli usando il metodo diretto (REJ) o la regressione logistica multinomiale (REG). Varie soglie decisionali sono state utilizzate (vedi materiali e metodi). VP: veri positivi; FP: falsi positivi.

MODELLO	SOGLIA	REJ		REG	
		VP	FP	VP	FP
M1	SR	0.97	0.23	0.91	0.09
	0.5	0.91	0.10	0.82	0.06
	0.7	0.55	0.02	0.49	0.03
M2	SR	0.82	0.08	0.90	0.08
	0.5	0.72	0.04	0.84	0.04
	0.7	0.54	0.01	0.75	0.02
M3	SR	0.65	0.12	0.69	0.13
	0.5	0.48	0.05	0.57	0.06
	0.7	0.14	0.00	0.24	0.01
M4	SR	0.38	0.20	0.52	0.08
	0.5	0.26	0.04	0.45	0.04
	0.7	0.01	0.00	0.08	0.01
M5	SR	0.46	0.19	0.55	0.10
	0.5	0.30	0.05	0.43	0.05
	0.7	0.00	0.00	0.04	0.01

La stima dei parametri demografici

In Tabella4.6 sono riportate le stime, e i relativi intervalli di credibilità, delle distribuzioni a posteriori di ogni parametro appartenente al modello M5 e in Figura4.2 e Figura4.3 sono rappresentate graficamente in confronto con la loro distribuzione a priori. Le dimensioni effettive moderne (N_1, N_2 e N_3) sono ben stimate come confermato dai valori di R^2 superiori al 60% e dalla

presenza di picchi nelle distribuzioni a posteriori (Tabella4.6 e Figura4.2). *C. rastrospinosus* e *C. hamatus* mostrano una dimensione effettiva molto simile, pari a circa 11 000 e 9 500 individui aploidi rispettivamente (usando come stimatore la mediana) e intervalli di credibilità che sebbene siano piuttosto larghi, non supportano comunque dimensioni effettive inferiori ai 2 500 – 3 500 individui (Tabella4.6). La dimensione effettiva di *C. myersi* è la più piccola, con una stima di circa 6 000 individui (HPD 732 – 15 326), cioè circa la metà rispetto alle altre due specie. Le dimensioni di popolazione antiche (N_{A1}, N_{A2}) non sono stimate correttamente, come suggerito dai valori estremamente bassi di R^2 e dall'ispezione visiva della forma delle distribuzioni a posteriori. Sebbene il modello con una migrazione Eemiana seguita da una migrazione Olocenica (M5) sia chiaramente preferito rispetto al modello con solo una migrazione Olocenica (M4), non sembra esserci abbastanza informazione nei dati per poter stimare i parametri della migrazione più antica. Infatti, i parametri m_{E12}, m_{E13} e m_{E23}, hanno le distribuzioni a posteriori molto sovrapposte con le distribuzioni a priori (Figura4.3) e valori di R^2 bassi (Tabella4.6). I tassi di migrazione moderni invece mostrano un'elevata correlazione con le statistiche descrittive e perciò sono stati combinati con le dimensioni effettive moderne, che mostrano i secondi valori più elevati di R^2, per massimizzare la capacità d'inferenza sugli eventi migratori recenti. Le tre dimensioni moderne (N_1, N_2 e N_3) e i tre tassi di migrazione Olocenica (m_{H12}, m_{H13} e m_{H23}) sono stati utilizzati per calcolare tre parametri congiunti, $m_{Hij(Ni+Nj)}$, corrispondenti al numero di migranti scambiati tra ogni coppia di specie per generazione (vedi Tabella4.6). In accordo con quanto indicato dai valori di distanza genetica tra le specie calcolati sui dati reali (Tabella4.3), *C. myersi* e *C. hamatus* mostrano il più basso livello di individui scambiati. Considerando la mediana della distribuzione a posteriori, il numero di migranti per generazione è di 1.95 individui diploidi (HPD 0 – 13.9). Valori più elevati sono stati stimati negli altri confronti, in particolare per la coppia *C. myersi* e *C. rastrospinosus*, dove il numero di individui scambiati per generazione è 11 (HPD 0 – 55.8). Le due specie più vicine, *C. rastrospinosus* e *C. hamatus*, mostrano livelli intermedi di migrazione con circa 6.9 (HPD 0 – 42.55) individui scambiati per generazione (Tabella4.6). Questi risultati assoluti devono essere interpretati con cautela, ma il rapporto relativo livelli di migrazione tra le coppie di specie è confermato dagli altri indici di tendenza centrale (Media e Moda) e anche dall'analisi dei tassi di migrazione non combinati. La distribuzione a posteriori della media del tasso di mutazione tra i loci (μ) è ampiamente sovrapposta con la distribuzione a priori, ma il valore di R2 associato ($R^2(\mu)=0.26$) conferma che dovremmo essere in grado di stimare il parametro. La stima puntuale di μ (mediana) è pari a 3.73×10^{-4}, leggermente inferiore alla mediana della distribuzione a priori (5×10^{-4}) ma perfettamente compatibile con essa. Questo risultato indica che la distribuzione a priori del tasso medio di mutazione sembra descrivere correttamente il pattern di variabilità genetica delle

specie, producendo stime a posteriori che si discostano di poco da quanto definito. Il parametro P, invece, non sembra essere stimato correttamente, avendo sia una distribuzione a posteriori senza un chiaro picco, sia un valore di R^2 estremamente basso ($R^2(P)= 3.29x10^{-5}$).

Tabella4.7: Stime dei parametri del modello M5. Tre indici di tendenza centrale e l'intervallo di credibilità (HPD) al 95% sono riportati in tabella. R2: proporzione della varianza dei parametri spiegata dalle statistiche descrittive.

Parametro		Media	Mediana	Moda	95%HPD min	95%HPD max	R^2
N_1	Cromosomi	13 216	11 222	6 722	1 464	30 653	0.63
N_2	Cromosomi	22 167	22 087	21 402	7 717	39 486	0.63
N_3	Cromosomi	19 534	18 719	14 275	4 770	36 776	0.62
N_{A1}	Cromosomi	19 563	19 386	500	500	32 901	$2.90x10^{-4}$*
N_{A2}	Cromosomi	17 547	16 030	500	500	30 171	$1.36x10^{-3}$*
μ	Tasso	$3.82x10^{-4}$	$3.73x10^{-4}$	$3.44x10^{-4}$	$1.66x10^{-4}$	$6.19x10^{-4}$	0.26
P	Tasso	0.36	0.34	0.11	0.00	0.75	$3.29x10^{-5}$*
m_{E12}	Tasso	$8.09x10^{-4}$	$4.28x10^{-5}$	$1.00x10^{-8}$	$1.00x10^{-8}$	$5.38x10^{-3}$	0.02*
m_{E13}	Tasso	$1.07x10^{-3}$	$5.51x10^{-5}$	$1.00x10^{-8}$	$1.00x10^{-8}$	$6.58x10^{-3}$	0.02*
m_{E23}	Tasso	$3.48x10^{-4}$	$4.23x10^{-6}$	$1.00x10^{-8}$	$1.00x10^{-8}$	$3.12x10^{-3}$	0.02*
m_{H12}	Tasso	$1.30x10^{-3}$	$7.04x10^{-4}$	$3.55x10^{-4}$	$1.00x10^{-8}$	$4.74x10^{-3}$	0.73
m_{H13}	Tasso	$2.02x10^{-4}$	$7.29x10^{-5}$	$1.00x10^{-8}$	$1.00x10^{-8}$	$6.62x10^{-4}$	0.73
m_{H23}	Tasso	$7.97x10^{-4}$	$3.82x10^{-4}$	$1.00x10^{-8}$	$1.00x10^{-8}$	$3.00x10^{-3}$	0.73
$m_{H12}(N_1+N_2)$	Cromosomi	35.6	22.0	14.4	$1.00x10^{-5}$	111.6	0.74
$m_{H13}(N_1+N_3)$	Cromosomi	7.7	3.9	$1.00x10^{-5}$	$1.00x10^{-5}$	27.8	0.74
$m_{H23}(N_2+N_3)$	Cromosomi	25.6	13.8	$1.00x10^{-5}$	$1.00x10^{-5}$	85.1	0.74

***i parametri che mostrano valori di R^2 minori di 0.1 sono stimati debolmente.**

4.4 DISCUSSIONE

La storia demografica del genere Chionodraco

Riscostruire la storia demografica delle popolazioni, utilizzando le informazioni presenti a livello del DNA, è un aspetto fondamentale per capire come eventi naturali o artificiali accaduti nel passato abbiano influenzato le caratteristiche genetiche osservabili nel presente. A questo scopo, l'Approximated Bayesian Computation si è dimostrato uno strumento estremamente utile per studiare la storia delle popolazioni attraverso l'analisi di modelli demografici complessi. In questo studio di biologia evoluzionistica, eseguito in collaborazione con l'Università di Padova, è stato studiato l'impatto dei due ultimi cicli glaciali sul genere *Chionodraco*, fornendo evidenze di introgressione tra specie di pesci antartici, e aprendo nuove prospettive nello studio della radiazione evolutiva di questo gruppo.

L'ipotesi di introgressione tra le specie è fortemente supportata dall'analisi ABC, dove tutti i modelli con migrazione interspecifica ottengono probabilità più alte rispetto al modello di isolamento, indicando quindi l'assenza di un isolamento riproduttivo completo. Questa evidenza è sottolineata inoltre dall'estrema concordanza tra i metodi di scelta del modello nell'escludere l'assenza d'ibridazione tra le specie, assegnando a questo scenario evolutivo una probabilità prossima a 0 anche con i metodi di scelta più sofisticati come la regressione logistica corretta (R-ADJ). Questo risultato sembra suggerire dunque che il processo di speciazione sia sostenuto da selezione ecologica divergente piuttosto che da un'incompatibilità degli ibridi (Stelkens et al. 2010). Anche lo scenario dove la migrazione è stata costante nel tempo non sembra essere particolarmente supportato, ottenendo probabilità più basse rispetto ai due modelli dove è presente migrazione o nel periodo interglaciale Olocenico o nei due periodi Eemiano e Olocenico. In accordo con le attese, eventi di ibridazione recenti sembrano dunque aver lasciato traccia nel pattern di variabilità ai loci microsatelliti studiati, sebbene anche fenomeni migratori antichi abbiano giocato un ruolo chiave.

Il modello con due eventi di migratori discontinui avvenuti durante i periodi interglaciali è il modello favorito tra tutti i modelli con migrazione interspecifica, utilizzando i metodi di scelta più credibili (REG e R-ADJ). Inoltre, le simulazioni condotte sul potere del metodo, indicano chiaramente che l'ABC ha l'abilità di discriminare tra i modelli soprattutto quando la regressione logistica viene impiegata. In accordo con il modello, durante i periodi glaciali, l'occupazione degli habitat da parte dei ghiacci avrebbe favorito l'isolamento in particolari regioni libere dai ghiacci (Thatje et al. 2005), con una possibile speciazione allopatrica come conseguenza. Durante i periodi interglaciali invece, la nuova disponibilità degli habitat dovuta al ritiro dei ghiacci, avrebbe

facilitato il contatto tra le specie con la conseguente ibridazione che oggi osserviamo (Near et al. 2004; Near et al. 2011). In questo contesto, i risultati ottenuti sottolineano il fatto che le condizioni climatiche attuali, limitando le possibilità di isolamento tra le specie, possono promuovere l'ibridazione interspecifica e perciò portare ad una perdita della diversità che oggi osserviamo. In questa prospettiva, si può pensare che la mono-specificità della gran parte dei generi osservati tra gli *Channichthyidi* sia dovuta al fatto che poche linee evolutive siano sopravvissute all'effetto di omogeneizzazione genetica durante i periodi interglaciali. In generale, questi risultati sembrano indicare che il riscaldamento globale possa incrementare il rischio di perdere in futuro parte dell'eccezionale radiazione evolutiva avvenuta in passato.

Performance dell'inferenza con ABC

Una fase importante dell'inferenza tramite Approximate Bayesian Computation è quella di valutare le prestazioni della metodologia nel riuscire a discriminare tra i diversi scenari evolutivi, e nella stima dei parametri, in modo da avere una misura quantitativa della credibilità dei risultati. Questo studio dimostra che l'ABC è uno strumento molto potente non solo per scegliere il modello migliore all'interno di un insieme di scenari possibili, ma anche per ottenere stime di parametri demografici utili a descrivere l'evoluzione delle specie con i relativi intervalli di credibilità. L'analisi di potenza ha dimostrato che con un numero non particolarmente elevato di loci microsatelliti è possibile identificare il modello di evoluzione più probabile con basse probabilità di commettere errori. In accordo con quanto riscontrato da Beaumont (2008), utilizzare un metodo di stima più raffinato come la regressione logistica multinomiale porta un guadagno in termini di performance, e minimizza la probabilità di commettere una falsa inferenza rispetto al metodo diretto. Le simulazioni impiegate in questa fase possono essere riutilizzate per aggiustare le probabilità calcolate con la regressione logistica, e calcolare la probabilità che ogni modello sia quello corretto. Questo procedimento ha il vantaggio di integrare l'informazione dell'analisi di potenza direttamente all'interno delle probabilità a posteriori di ogni modello e renderne più facile l'interpretazione. Per quanto riguarda la stima dei parametri demografici, le dimensioni effettive delle popolazioni moderne e i tassi di migrazione (e la combinazione dei due) hanno il potenziale per essere stimati correttamente, dato che le statistiche descrittive spiegano una proporzione sostanziale della loro varianza e le distribuzioni a posteriori hanno picchi di probabilità per valori dei parametri diversi rispetto alle distribuzioni a priori. Uno dei limiti principali dell'ABC è che non è in grado di "stimare" il modello evolutivo migliore in assoluto per le popolazioni, ma solamente di trovare il modello più supportato tra quelli definiti dall'utente. Di conseguenza, se all'interno dell'insieme dei modelli da testare non è presente il modello vero, l'ABC comunque tenderà a

favorire il modello che più si avvicina a riprodurre i dati osservati, anche se questo non è il reale. Questo limite viene in parte superato eseguendo una serie di test per verificare la bontà di uno scenario demografico, come ad esempio verificare che le simulazioni prodotte dal modello maggiormente supportato includano nel loro range di variazione le statistiche descrittive osservate nei dati reali, oppure eseguendo "test predittivi a posteriori" basati sull'idea che se il modello selezionato dall'ABC e i parametri stimati sono corretti, si dovrebbe riuscire a riprodurre le statistiche descrittive osservate nei dati reali eseguendo delle simulazioni a posteriori (Gelman et al. 2003). Infine, la selezione delle statistiche descrittive potrebbe aver avuto un effetto sui risultati. Questa fase, ad oggi, è uno dei più attivi ambiti di ricerca applicata all'ABC perché fondamentale per una corretta inferenza, ma sebbene diversi metodi siano stati pubblicati per affrontare questo problema, ancora non si è raggiunta un'efficienza computazionale sufficiente. Ad esempio, il metodo più rigoroso a oggi disponibile, proposto da Joyce e Marjoram (2008), prevede di calcolare un indice di qualità per ogni combinazione di statistiche descrittive presenti in un insieme iniziale. Applicato alla nostra situazione, partendo da un insieme di 29 elementi, si dovrebbero testare più di $5x10^8$ combinazioni diverse rendendo di fatto proibitivo questo tipo di analisi. Tuttavia, la bontà delle statistiche impiegate è confermata dai risultati ottenuti in altri studi dove sono state utilizzate con successo (vedi ad esempio Guillemaud et al. 2010 per le statistiche intra popolazione e Ghirotto et al. 2011 per l'uso delle categorie di alleli privati, condivisi e F_{ST}).

Sebbene lo studio dei dati genetici del genere *Chionodraco* abbia fornito numerose informazioni sulla sua dinamica d'ibridazione, la procedura inferenziale con ABC può essere certamente migliorata. L'incremento del numero di loci microsatelliti del campione dovrebbe riflettersi in aumento della qualità della stima come dimostrato in studi precedenti (Excoffier et al. 2005), portando a stime più precise e a intervalli di credibilità più stretti. Inoltre, con la giusta quantità d'informazione genetica dovrebbe essere possibile stimare direttamente dai dati il tempo e la durata degli eventi d'ibridazione congiuntamente alla loro intensità e verificare se i tempi stimati correlano con eventuali eventi climatici. Aggiungere un diverso tipo di marcatore genetico dovrebbe aumentare le performance dell'analisi. I marcatori microsatelliti sono caratterizzati da un elevato tasso mutazionale, che li rende un tipo di marcatore molto informativo riguardo a fenomeni demografici recenti, ma molto limitato nel fornire informazioni su fenomeni molto antichi come la divergenza tra le specie. Nei modelli demografici analizzati in questo studio, questo problema è stato risolto fissando i tempi di divergenza secondo le stime ottenute su un marcatore mitocondriale. Includere direttamente nell'analisi, oltre ai microsatelliti, anche sequenze nucleari, permetterebbe di avere il giusto bilanciamento nel potere inferenziale di fenomeni antichi e recenti, lasciando che siano i dati genetici a supportare i parametri demografici più adatti. In conclusione, l'ABC si è

dimostrato un metodo molto potente per identificare ibridazione interspecifica nel genere *Chionodraco*, anche se ulteriori analisi, specialmente includendo altri tipi di marcatori genetici, sembrano essere necessarie per una caratterizzazione più precisa del modo e dei tempi in cui i fenomeni migratori siano avvenuti e di come siano stati influenzati da eventi climatici passati.

4.5 MATERIALI SUPPLEMENTARI

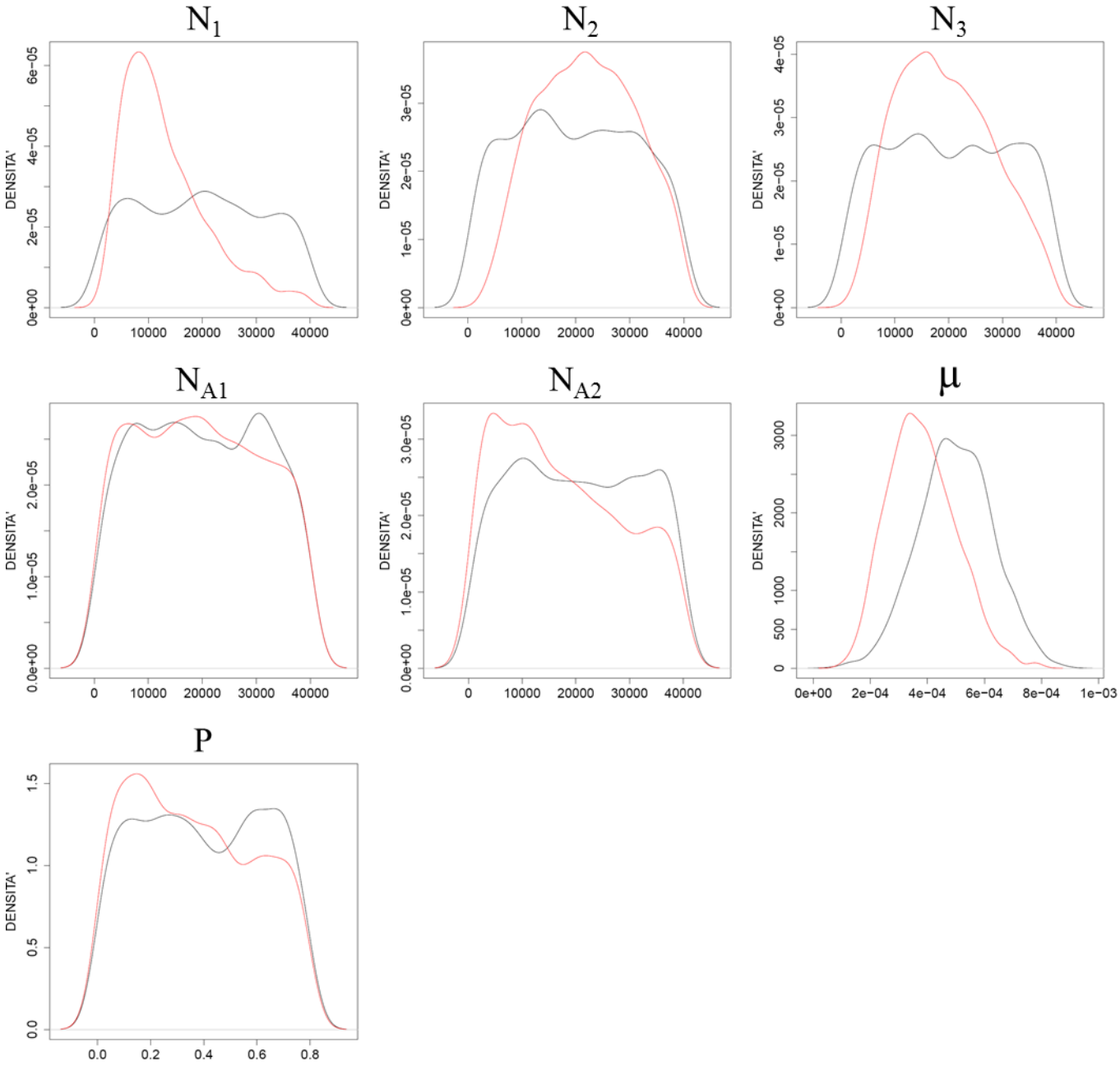

Figura4.2: Distribuzioni a posteriori (n rosso) delle dimensioni effettive (N_1, N_2, N_3) e dei parametri mutazionali (μ, P) del modello M5. La curva nera rappresenta la distribuzione a priori.

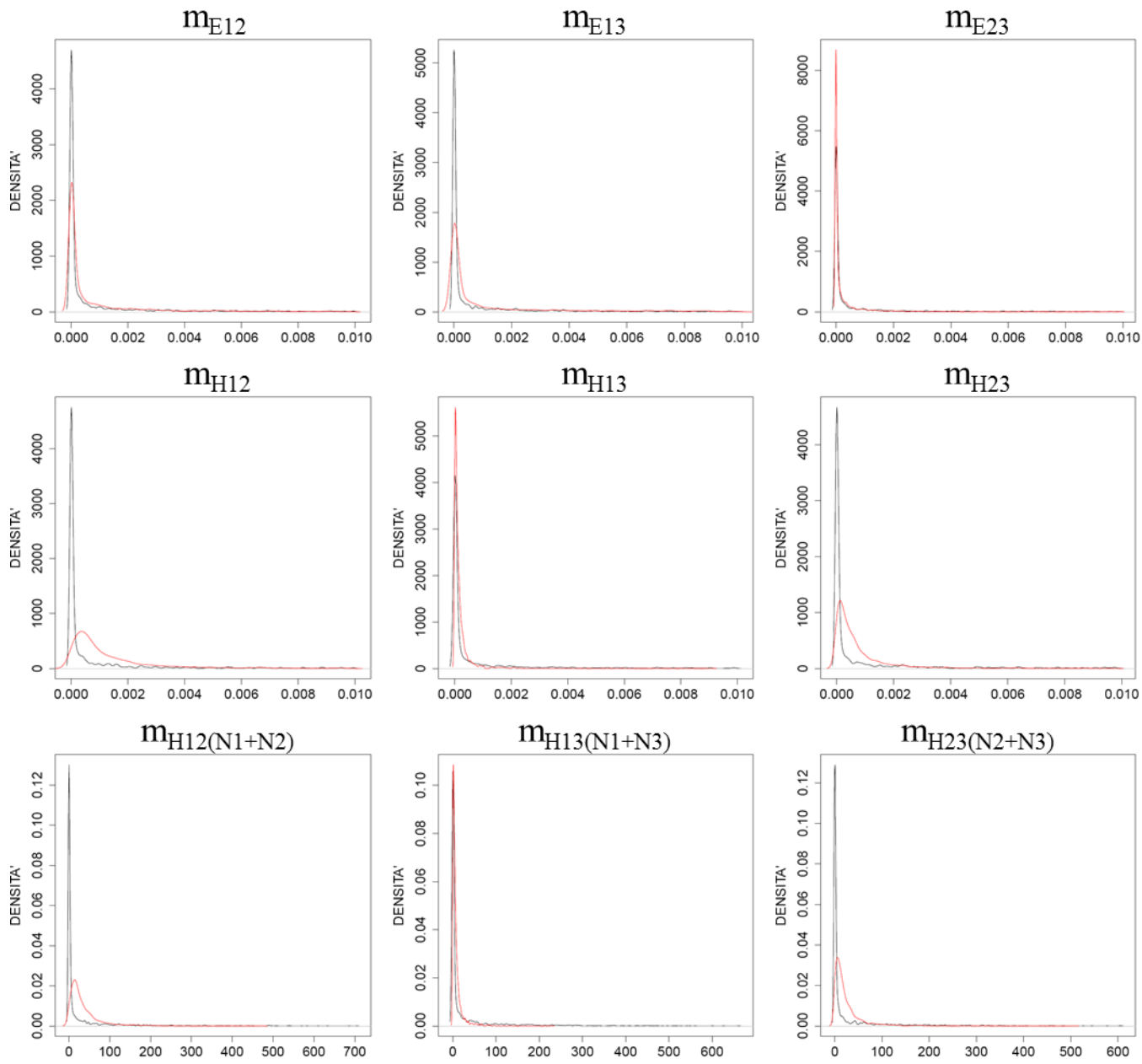

Figura4.3: Distribuzioni a posteriori (in rosso) dei parametri relativi alla migrazione del modello M5. La curva nera rappresenta la distribuzione a priori.

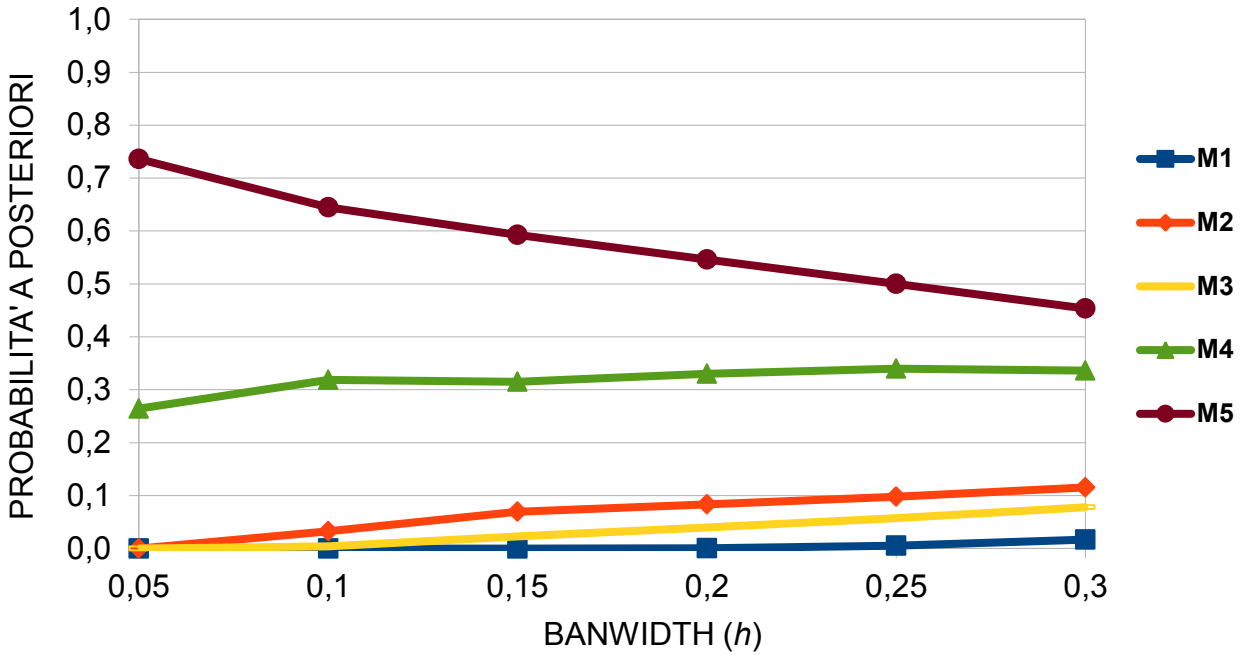

Figura4.4: dipendenza della probabilità a posteriori dei modelli dal valore bandwidth (*h*) utilizzato secondo l'approccio R-ADJ. I diversi colori corrispondono ai modelli considerati.

CONCLUSIONI

In questa tesi ho applicato diversi approcci, basati sull'utilizzo delle simulazioni genetiche, per rispondere a domande differenti di genetica di popolazioni. Le simulazioni si sono dimostrate uno strumento molto flessibile e veloce per studiare un sistema, e capaci di estrarre informazioni impossibili da ottenere con altri approcci. Nel primo studio presentato, l'obiettivo è stato quello di studiare il comportamento di un metodo per l'inferenza demografica quando una propria assunzione specifica non era rispettata. In questo caso, attraverso l'uso delle simulazioni coalescenti è stato possibile generare migliaia di dati genetici secondo un modello demografico dove la condizione di isolamento è stata violata. Lo studio ha identificato quali sono gli effetti dei fenomeni migratori sulle ricostruzioni demografiche, basate sullo skyline plot, e in che condizioni questi si fanno più marcati. Questi risultati sono stati ottenuti grazie a un controllo preciso dei parametri del modello e che ha permesso di studiare nel dettaglio l'influenza della violazione sulle stime demografiche.

Il merito di riuscire a generare dati genetici che riflettano in modo accurato la demografia delle popolazioni è certamente del modello Coalescente (Kingman 1982ab). Il successo di questo modello deriva dal fatto che riesce a riprodurre la variabilità genetica in un campione individui, data una certa demografia, senza dover simulare tutta la popolazione. Dalla sua prima formulazione nel 1982, il modello è stato successivamente esteso sia nella parte demografica, permettendo di gestire popolazioni multiple, popolazioni strutturate, selezione, e cambiamenti demografici, sia nella parte genetica, introducendo differenti tipi di marcatori genetici, modelli mutazionali e ricombinazione (vedi la review Hoban et al per maggiori dettagli). Il vantaggio di poter modellare con precisione la storia di una popolazione, e ottenere da essa le informazioni genetiche in modo efficiente, ha promosso il rapido sviluppo di strumenti software basati sul Coalescente che ad oggi conta decine di implementazioni diverse. In questo contesto, l'ampia gamma di simulatori permette di testare la robustezza di un qualsiasi metodo di indagine basato sul DNA in moltissime condizioni di potenziale interesse.

L'uso di un approccio simulativo è stato applicato in questa tesi anche per rispondere ad un problema legato alla scelta della strategia di campionamento migliore per massimizzare l'inferenza demografica. In questo caso, il problema è stato quello di identificare il migliore modo di campionare un centro numero di sequenze nel tempo, in modo da massimizzare la probabilità di ricostruire la giusta dinamica demografica di una popolazione. Questo problema ha notevoli ricadute pratiche in quanto ogni campione che non sia preso in una popolazione moderna deve essere tipizzato geneticamente da reperti museali, facendone lievitare i costi per la produzione del

dato genetico in laboratorio. Attraverso l'uso delle simulazioni è stato possibile creare dataset genetici con differenti organizzazioni temporali dei dati al loro interno, per ognuno di loro calcolare diversi indici di qualità della ricostruzione, e infine identificare il migliore. I risultati dell'analisi delle prestazioni in uno scenario di riduzione demografica hanno selezionato come schema migliore la combinazione di un campione moderno e di uno antico, preso prima della riduzione demografica. Questi risultati, ottenuti completamente con dati genetici generati al computer, forniscono un consiglio su com'è meglio organizzare la fase di campionamento se si ipotizza che la popolazione che si intende studiare abbia caratteristiche genetiche (tipo di marcatore, livello di polimorfismo genetico, ecc…) e demografiche (intensità e tempo della riduzione) simili alle condizioni prese in considerazione nello studio simulativo. Nei casi in cui i risultati non siano estendibili, come ad esempio lo studio di una popolazione in espansione demografica, questo studio fornisce le linee guida da seguire per replicare l'esperimento ad un problema simile. Grazie alle crescenti capacità di calcolo dei computer e dei cluster di calcolo, sempre più presenti nei centri di ricerca, le simulazioni si candidano come uno strumento utile nella fase di progettazione di un esperimento scientifico, fornendo indicazioni utili a massimizzare l'inferenza statistica e a risparmiare risorse economiche.

Uno degli svantaggi dell'utilizzare le simulazioni basate sul Coalescente negli studi di simulazione deriva dalla sua natura stocastica: è molto improbabile, infatti, che due simulazioni dello stesso processo demografico diano un pattern di variabilità genetica identico. Il Coalescente permette, infatti, di derivare il livello atteso di polimorfismo (e relativa varianza) usando particolari parametri demografici, ma non di conoscerne a priori la quantità esatta nella singola simulazione. A livello pratico, questo comporta l'analisi ripetuta di numerose simulazioni provenienti dalle stesse condizioni demografiche (chiamate anche repliche) per tenere conto della variabilità interna dei dati generati. Normalmente, il numero di repliche varia dalle centinaia alle migliaia ma dipendente soprattutto dalla tipologia del dato genetico da simulare e dal livello di precisione voluto dall'utente. Inoltre, il numero di parametri demografici che si vogliono analizzare influenza pesantemente il numero totale di simulazioni da eseguire. Più si vuole un'analisi dettagliata, più combinazioni di parametri demografici devono essere testate, portando generalmente la complessità del problema a un ordine $O(k^n)$, dove k è peri al numero di differenti combinazioni di parametri e n è il numero di repliche per ogni combinazione. Di conseguenza, in uno studio di simulazione si è spesso alle prese con migliaia o centinaia di migliaia di dataset genetici da analizzare e questo può essere un fattore limitante nei casi in cui non si disponga delle capacità elaborative necessarie.

Inoltre, il tempo necessario alla produzione delle simulazioni è direttamente proporzionale al numero e alla complessità delle regioni genomiche che s'intende riprodurre, oltre che al tipo di

simulatore impiegato. I simulatori in forward, permettono un'accurata descrizione dei processi demografici, simulando la dinamica di ogni singolo individuo della popolazione nel tempo. Attraverso il loro utilizzo è possibile generare singoli cromosomi o addirittura l'intero genoma di un individuo (Carvajal-Rodriguez 2008; Padhukasahasram et al. 2008) secondo scenari demografici estremamente complessi, differenti pressioni selettive e modi di accoppiamento degli individui. Questo tipo di simulatori permettono di ottenere dati genomici pressoché per qualsiasi condizione che si voglia indagare, ma attualmente non possiedono quell'efficienza tale, in termini di tempo, da farli diventare il tipo di simulatori predominante. Al contrario, i simulatori in backward possiedono questo tipo di efficienza, soprattutto dopo l'introduzione del "sequential markov coalescent" (McVean e Cardin 2005; Marjoram e Wall 2006). Grazie a questa particolare approssimazione del Coalescente (gli eventi di ricombinazione sono modellizzati in modo più rapido), è possibile simulare larghe regioni genomiche, con ricombinazione, secondo un qualsiasi scenario demografico. Nonostante questi miglioramenti, si è ancora lontani dall'obiettivo di simulare, in maniera efficiente, dati di polimorfismo genetico lungo tutto genoma simili a quanto è attualmente possibile produrre con le nuove tecnologie di sequenziamento su larga scala (Dudek et al. 2006). Inoltre, nel prossimo futuro, per poter utilizzare le informazioni provenienti da simulazioni di dati genetici di nuova generazione, sarà necessario integrare nei simulatori la gestione degli errori di tipizzazione genetica di cui queste tecnologie di sequenziamento sono caratterizzate.

Il miglioramento delle prestazioni dei simulatori in backward ha di fatto reso possibile lo sviluppo di tecniche di inferenza statistica basate sulle simulazioni come l'Approximated Bayesian Computation. In questo caso, milioni di simulazioni devono essere prodotte in tempi accettabili per poter raggiungere il livello di precisione richiesto. Lo svantaggio in termini di tempo computazionale viene ricompensato dalla possibilità di analizzare modelli complessi non trattabili con i metodi canonici, di fatto rendendo questa tecnica una delle più promettenti in genetica di popolazione. In questa tesi, questa metodologia statistica è stata applicata ad uno studio di biologia evoluzionistica, per studiare gli eventi di ibridazione tra tre specie di pesci antartici appartenenti al genere *Chionodraco* e ha permesso di identificare gli ultimi due cicli glaciali come principali responsabili dell'introgressione avvenuta. Il numero totale di simulazioni necessarie per portare a compimento l'analisi è stato di poco superiore ai 5 milioni, e se si considera che ad ogni ciclo venivano prodotti 8 loci microsatelliti, il numero totale di locus genetici simulati è stato a circa 40 milioni per un totale di circa 2 000 ore/processore. Da questi numeri si deduce che, l'incremento delle capacità di calcolo, deve essere seguito da un miglioramento dell'efficienza degli algoritmi di simulazione, in modo da facilitare la diffusione di questa metodologia statistica particolarmente utile per analizzare dati reali con modelli sempre più realistici. Ad esempio, Neuenschwander et al.

(2008) hanno applicato un tipo di analisi ibrida forward-backward per studiare la ricolonizzazione postglaciale del bacino del fiume Reno in Svizzera da parte del *Cottus gobio*. Secondo il loro approccio, in una prima fase centinaia di migliaia di eventi di ricolonizzazione sono stati simulati in forward tenendo in considerazione le informazioni spaziali e geografiche della zona e, in una fase successiva, è stato utilizzato un simulatore in backward per generare la variabilità genetica in ognuna delle demografie generate in precedenza. In seguito, l'ABC è stato utilizzato per stimare i parametri della colonizzazione più verosimili date le informazioni genetiche campionate nelle popolazioni moderne. Questa combinazione di diverse tecniche di simulazione, si è dimostrata un approccio molto potente per studiare nel dettaglio un evento complesso come può essere la colonizzazione di un nuovo habitat, portando alla stima di parametri chiave come il tempo della colonizzazione completa, le dimensioni effettive delle popolazioni coinvolte e i tassi di migrazione. Tuttavia, l'analisi di soli sei loci microsatelliti ha richiesto più di 15 000 ore di calcolo, un tempo sproporzionato in relazione ai dati genetici impiegati, sottolineando la difficoltà dei simulatori, e delle tecniche inferenziali attuali, nel gestire livelli troppo alti di realismo.

In conclusione, le simulazioni si sono dimostrate uno strumento estremamente utile per studiare il comportamento di metodi statistici, per trovare strategie di campionamento ottimali e per inferire la storia evolutiva di specie animali. Sebbene in molti casi le simulazioni siano estremamente costose dal punto di vista computazionale, hanno le potenzialità per diventare uno strumento standard per l'analisi dei dati di variabilità genetica presenti e futuri.

ALLEGATO

MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 2609–2625 doi: 10.1111/j.1365-294X.2010.04690.x

INVITED REVIEW

ABC as a flexible framework to estimate demography over space and time: some cons, many pros

G. BERTORELLE,* A. BENAZZO* and S. MONA*†‡

*Department of Biology and Evolution, University of Ferrara, Via Borsari 46, 44100 Ferrara, Italy, †CMPG, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland, ‡Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Abstract

The analysis of genetic variation to estimate demographic and historical parameters and to quantitatively compare alternative scenarios recently gained a powerful and flexible approach: the Approximate Bayesian Computation (ABC). The likelihood functions does not need to be theoretically specified, but posterior distributions can be approximated by simulation even assuming very complex population models including both natural and human-induced processes. Prior information can be easily incorporated and the quality of the results can be analysed with rather limited additional effort. ABC is not a statistical analysis *per se*, but rather a statistical framework and any specific application is a sort of hybrid between a simulation and a data-analysis study. Complete software packages performing the necessary steps under a set of models and for specific genetic markers are already available, but the flexibility of the method is better exploited combining different programs. Many questions relevant in ecology can be addressed using ABC, but adequate amount of time should be dedicated to decide among alternative options and to evaluate the results. In this paper we will describe and critically comment on the different steps of an ABC analysis, analyse some of the published applications of ABC and provide user guidelines.

Keywords: approximate Bayesian computation, likelihood-free inference, molecular ecology, population demography, population genetics, population history

Received 19 January 2010; revision received 20 April 2010; accepted 21 April 2010

Introduction

Population genetics is the analysis and understanding of genetic variation within and between populations. Early population geneticists, possibly also as a consequence of the paucity of empirical data, were mainly concerned with the theoretical framework of this discipline. Assuming simple demographic and evolutionary models, expected genetic variation patterns were theoretically predicted and sometimes compared with the available genetic information. During an intermediate phase from the 1970s to the early 1990s, when classical genetic markers were easily typed and the use of molecular markers began to spread following the introduction of the PCR, descriptive analyses of genetic variation dominated. Methods such principal component analysis (PCA), spatial autocorrelation, and analysis of molecular variance (AMOVA) were widely used to describe patterns and informally compare hypotheses (e.g. Menozzi *et al.* 1978; Sokal *et al.* 1987; Excoffier *et al.* 1992). Parameter estimation and probability-based comparison of different scenarios were limited and imprecise, due to the fact that contemporary models were unrealistic and that more complex demographic and genetic models were theoretically intractable or computationally prohibitive. More recently, the increased speed and power of personal computers favoured the spread of Monte Carlo algorithms. Likelihood functions can be approximated thanks to Markov Chain Monte Carlo (MCMC) methods

Correspondence: Giorgio Bertorelle, Fax: +390532249771; E-mail: ggb@unife.it

(e.g. Kuhner *et al.* 1995; Nielsen & Wakeley 2001; Drummond *et al.* 2002) and brute power can be used to simulate gene genealogies under virtually any demographic and genetic model and to approximate the likelihood functions even without explicitly defining them (e.g. Fu & Li 1997; Tavare *et al.* 1997; Beaumont *et al.* 2002). This latter approach, called approximate Bayesian computation (ABC) in its Bayesian version, is the topic of this review. We believe that ABC is matching, for the first time in population genetics studies, abundant genetic data and realistic (which usually means complex) evolutionary scenarios, allowing (i) the simultaneous estimation of posterior distributions for many parameters relevant in ecological studies; (ii) the probabilistic comparison of alternative models; and (iii) the quantitative evaluation of the results' credibility.

Approximate Bayesian computation is intuitively very easy: millions of genealogies are simulated assuming different parameter values and under different models and the simulations that produce genetic variation patterns close to the observed data are retained and analysed in detail. Parameter values and model features in the retained simulations are of course interesting since they are able to generate data sets with some properties, measured by summary statistics (SuSt hereafter), found in the observed data. At the same time, even if software packages are now available (e.g. Cornuet *et al.* 2008; Wegmann *et al.* 2010), ABC is not (yet?) user-friendly. Users are typically required to: (i) carefully consider each step in the ABC protocol since consensus on the best way to proceed has not been reached; and (ii) estimate the quality of the final results. In short, ABC is mathematically graceless and rather intricate to apply, but very flexible and powerful. In this review we will describe and critically comment on the different steps of an ABC analysis, analyse some of the published applications of ABC and provide throughout the paper some user guidelines. We will not discuss the recent criticisms to ABC and in general to Bayesian methods (Templeton 2010a,b). Detailed answers can be found, for example, in Beaumont *et al.* (2010).

First of all, we present the main ABC concepts in a historical perspective.

ABC: main concepts and history

Origins

The basic idea of ABC can be found in two papers published in February 1997. Stimulated by Templeton (1993) to find a correct estimator of the time to the most recent common ancestor (TMRCA) for a set of DNA sequences and assuming a simple demographic model of a single demographically stable population, Fu & Li (1997) and Tavaré *et al.* (1997) proposed simulating artificial data-sets and using SuSt to select among them. The selected data-sets, used to estimate the posterior distribution of the TMRCA, were either those having exactly the same maximum number of pairwise differences k_{max} as the observed data set (Fu & Li 1997) or those having a gene genealogy whose total length was compatible with the observed number of segregating sites, S (Tavare *et al.* 1997). The former approach can be almost considered 'theory-free', since knowledge of probability functions is not needed to approximate likelihood or posterior densities of the quantities of interest under any specified demographic and mutational model. This is the reason why the Fu & Li (1997) idea can, in principle, be applied to any demographic scenario, favouring its spread and extension with little theoretical effort. On the other hand, the algorithm proposed by Tavaré *et al.* (1997) had the merit of explicitly introducing the Bayesian component [the parameter $\theta = 4N\mu$ was not fixed as in Fu & Li (1997), but sampled from a prior distribution], which is a key aspect of modern ABC.

All the information contained in the data is not captured by a single SuSt. Also, if simulated data-sets are retained only when they show a SuSt identical to the SuSt observed in the real data, a large number of simulations are discarded. Weiss & Von Haessler (1998) addressed these two different but related problems suggesting that more SuSt should be used to better compute the distance between simulated and observed data sets and only the simulations in which the distance between simulated and observed data sets was higher than a specific threshold should be discarded. In particular, Weiss & Von Haessler (1998) used S and k as SuSt, where k is the mean pairwise difference between DNA sequences, and applied the distance threshold to k excluding the simulations where $|k' - k|$ was larger than 0.2 ($|\ |$ indicates the absolute value, and the presence or absence of the prime refers to the SuSt in the simulated and real the data sets, respectively). Weiss & Von Haessler (1998) also pioneered the use of simulations and SuSt to compare alternative demographic models, but did not incorporate, as was done a year later by Pritchard *et al.* (1999), the Bayesian step suggested by Tavaré *et al.* (1997).

In synthesis, the most important aspect of ABC which favoured its rapid development is that the likelihood function does not need to be specified. Using ABC, the posterior distribution of a parameter given the observed data, $P(\theta | D)$, can be empirically reconstructed since the likelihood is positively related to the distance between summary statistics computed in real and simulated data sets. More formally, when data are replaced by summary statistics, the reconstructed distribution is

$P(\theta \mid \rho(\mathrm{SuSt}_{sim}, \mathrm{SuSt}) \leq \varepsilon)$ (hereafter, $P(\theta \mid \rho \leq \varepsilon)$, where ρ is any distance metrics between observed and simulated SuSt and ε an arbitrary threshold. In the limit of $\varepsilon \to 0$ and if SuSt are sufficient (i.e. they capture all the relevant features of the data), $P(\theta \mid \rho \leq \varepsilon)$ will match exactly $P(\theta \mid D)$. The idea of ABC is that a good balance between accuracy and efficiency can be reached for small values of ε.

The formal definition of ABC

Beaumont *et al.* (2002) formalized and generalized the ABC approach. They introduced a series of improvements, evaluated the performance of ABC finding a reasonably good agreement with full-likelihood methods under some simple scenarios and discussed in some detail the challenging aspects associated with the choice of SuSt and of the most appropriate distance threshold ε. The actual birth of ABC coincides with this study.

The major improvement introduced by Beaumont *et al.* (2002) is the regression step. Roughly speaking, the slope of the regression line (regression is linear) between a parameter and the vector of SuSt, estimated using the retained simulations (regression is local) and giving more weight to the simulations producing SuSt closer to the observed values (regression is weighted), is used to modify the retained parameters' values and thus mimics a situation in which all simulations produce SuSt equal the observed values. If the chosen ε is very low, the regression step is unnecessary, but the acceptance rate will be very low and a very large numbers of simulations will be required in most cases. Increasing ε, the acceptance rate obviously increases, but in this case the regression step becomes important to improve the approximation of $P(\theta \mid \rho = 0)$ by $P(\theta \mid \rho < \varepsilon)$. For multiple SuSt, ρ is usually computed as the Euclidean distance between observed and simulated SuSt. The regression step aims specifically at reducing this discrepancy between simulated and observed SuSt by weighting and adjusting the parameters in the retained simulations, thus requiring fewer simulations. In these circumstances, Beaumont *et al.* (2002) showed that the regression method clearly outperforms the simple rejection algorithm, in which retained parameters are directly used to reconstruct their posterior distribution.

Recently, Leuenberger & Wegmann (2010) reformulated the regression step using the General Linear Model (GLM). SuSt are here response variables with explicit causes within the model, whereas the regression model introduced by Beaumont *et al.* (2002) considered the SuSt as explanatory variables. Some pros and cons of this approach are discussed in the 'Step 8' section. Under a simple one-population model which allows (for comparison) the analytical computation of the results, the ABC-GLM approach provide a good approximation of the posterior probability of the parameters [i.e. it produces $P(\theta \mid \rho < \varepsilon)$ close to $(P(\theta \mid D)]$, even when the chosen ε was moderately large (Leuenberger & Wegmann 2010).

ABC, MCMC and importance sampling

All simulations are independent under the ABC approach. This means that if a simulated genealogy produces an interesting data-set, i.e. a data-set with SuSt very similar to the observed values, the next simulation can be absolutely useless. In other words, approaching by chance the real values of the parameters during the simulations does not affect the machinery of the method. This sounds inefficient and Marjoram *et al.* (2003) introduced an algorithm to link simulations along a Markov chain path. The parameters for each new simulation are no longer sampled randomly from their prior distributions but are obtained starting from the values used in the previous simulation. The parameter space is explored as in classical MCMC methods, but a substantial difference is introduced. In the Metropolis–Hasting ratio, which is used to decide whether or not to accept a proposed parameter value, the likelihood term is replaced by an indicator function that takes a value of 1 if a simulated data set produces a distance between observed and simulated SuSt below ε and 0 otherwise. As expected, the acceptance ratio and thus the algorithm speed increase, but simulations are not independent any more. One practical advantage of ABC, that simulations for a single analysis can be run on many independent computers and simply pooled at the end, is therefore lost with the introduction of MCMC (but see Wegmann *et al.* 2009 for a possible solution). Embedding the ABC analysis in a MCMC setting raises new problems, some of which are common to any MCMC analysis (e.g. determining the length of the chain, monitoring its mixing and assessing the convergence) and some others are specific of ABC–MCMC. Among the latter, the choice of ε and the definition of the proposal distribution appear crucial to prevent the chain to stick to regions of low likelihood (Sisson *et al.* 2007). Bortot *et al.* (2007) proposed to augment the parameter space by treating ε as an additional parameter and Wegmann *et al.* (2009) introduced a preliminary simulation step to select the threshold ε and to set the proposal distribution.

Additional Monte Carlo schemes, such as population (Cappè *et al.* 2004) and sequential (Doucet *et al.* 2001) Monte Carlo, are under development. Here, importance sampling arguments in various flavours and with various acronyms (ABC-PRC, ABC-PMC, ABC-SMC) are used with the same purpose of MCMC settings to better

explore the parameter space, avoiding the simulation (and the analysis) of unrealistic scenarios (see e.g. Sisson *et al.* 2007; Beaumont *et al.* 2009; Toni *et al.* 2009). Preliminary simulations are used to identify a set of parameters vectors, called *particles*, which are within a certain distance ε from the observed data. The particles are then repeatedly re-sampled (according to a weighting scheme that considers the prior distributions), perturbed (using a transition kernel) and filtered (on the basis of new set of simulations and a decreased threshold ε). The particles after this iterative process tend to converge to a sample from the posterior distribution of the parameters. A final regression adjustment on the retained parameters can be easily applied to all these, as well as MCMC, algorithms (Beaumont *et al.* 2009; Wegmann *et al.* 2009; Leuenberger & Wegmann 2010).

The performances of ABC modified via MCMC or importance sampling have been analysed on simple simulated or real data sets, but the few results available appear controversial. For example, standard ABC, ABC-PMC (ABC with population Monte Carlo, Beaumont *et al.* 2009) and ABC-MCMC (under the Bortot *et al.* 2007 implementation) behave similarly when the computing times are kept identical [Fig. 2 in Beaumont *et al.*(2009)], whereas the ABC-MCMC implemented by Wegmann *et al.* (2009) seems to reach the performances of conventional ABC with a reduction of computational time.

ABC and model selection

Selecting among alternative models under the conventional ABC framework is, at least in principle, even simpler than parameter estimation. The mechanism of the direct method introduced by Weiss & Von Haessler (1998) and Pritchard *et al.* (1999) is straightforward. After pooling all the simulations generated by different models and retaining only those within a distance threshold from the real data, the posterior probabilities of each model is approximated by the fraction of simulations produced by each of them. Accuracy can be very low if the distance threshold ε is not close to 0, but can be improved using the logistic regression approach introduced by Beaumont (2008). The direct and the logistic approaches have been used and compared in various studies (Beaumont 2008; Cornuet *et al.* 2008; Guillemaud *et al.* 2010) and the possible advantages of some recent and more complex alternatives (see Toni *et al.* 2009; Leuenberger & Wegmann 2010) are under investigation.

ABC in nine steps

Here we update, extend and generalize the ABC scheme reported in Excoffier *et al.* (2005). The steps of a standard ABC analysis, which should be more technically defined as 'rejection ABC', are reported in Fig. 1. Running such ABC analysis rigorously requires careful development of each module, assemblage and validation.

Recently, two implementations of non-standard ABC (using MCMC and PMC) have become available for general users within the set of programs called ABCtoolbox (Wegmann *et al.* 2010). The use of these variants implies the replacement of a specific ABC module, but the general scheme and strategy for the whole analysis does not vary. We have therefore limited our description to the standard ABC.

Step 1: setting the scene

The *model*, i.e. the history and the demography of the populations with the associated parameters together with the genetic parameters relevant for the typed loci, needs to be clearly specified. Unsampled populations can and should be included in the model if they are potentially relevant for the sampled populations. In principle, the complexity of the scenario is not a limiting factor. Almost any demographic event, including migration, colonization, extinctions, divergence, population size changes, mass migrations or translocations, can be easily simulated and thus considered by ABC. Given this opportunity offered by ABC, it is easy to understand why classical population genetics models such as the stepping stone model (Kimura & Weiss 1964) or the divergence-with-isolation model (Wakeley & Hey 1997) appear unrealistic.

The parameters used to specify the model for an ABC analysis are the classic demographic and ecological parameters (e.g. population sizes, migration/growth/admixture rates, carrying capacities), the ages of any sort of natural or human-mediated population event (e.g. population split, translocation, invasion, bottleneck) and the genetic parameters (mutation and recombination rates with associated sub-parameters if needed). Under the hyperprior approach, particularly suitable for situations in which many loci and/or many species are simultaneously analysed (Excoffier *et al.* 2005; Hickerson *et al.* 2006; Beaumont 2008) the parameterization is hierarchical: hyper-parameters define some general feature (e.g. the mean mutation rate at a certain number of microsatellites) and single parameters are defined conditionally. In this way, the parameter space is explored more efficiently and more meaningfully. In principle, even aspects strictly related to the structure of the model, such as the size of river segments (see Neuenschwander *et al.* 2008) or the number of populations, can be defined as parameters to be estimated. This approach can be useful especially in

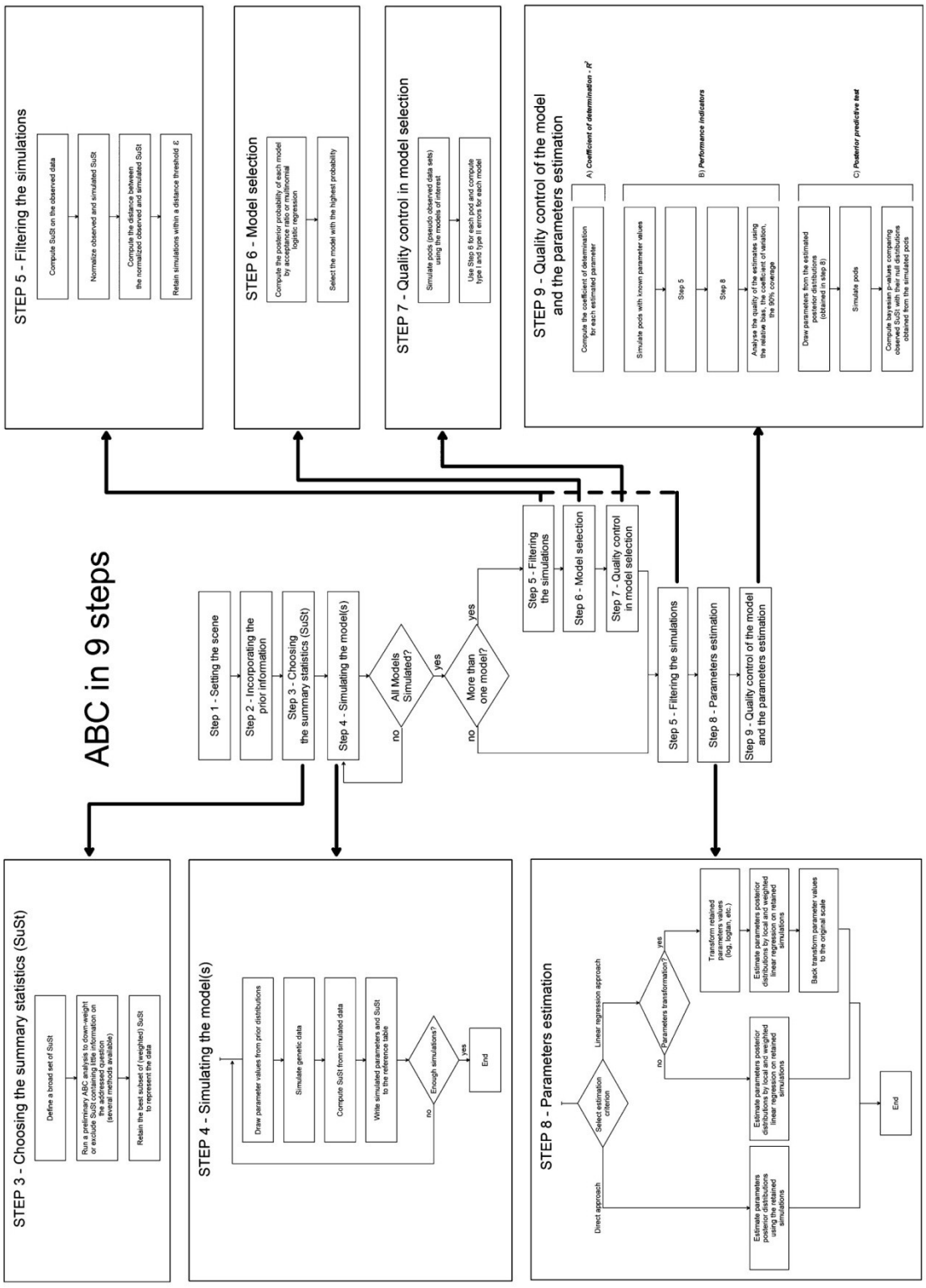

Fig. 1 ABC in nine steps.

preliminary ABC tests to identify and then fix, some aspects of the model.

Depending on the available data and on the relative influence of different events and parameters on the genetic variation pattern, increasing the complexity of the model can be either useless, time consuming or result in poor exploration of the parameter space. Some simple reasoning or preliminary simulations can be used to better understand the impact of different aspects of the model on the genetic variation pattern and, consequently, used to fix the value of some parameters and/or simplify the model (e.g. Estoup & Clegg 2003; Pascual *et al.* 2007; Ludwig *et al.* 2008). We believe, however, that ABC should be, at least at the beginning of the analysis, used at its maximum potential power, i.e. devising realistic models. A detailed analysis of the results, also comparing different runs (e.g. Hickerson *et al.* 2006), will help to determine their robustness.

Clearly, if different models are going to be compared, all of them need to be defined at the outset. Models can be nested or non-nested and it may be also interesting to compare the differences between the results provided by ABC under realistic and simplified versions of a model.

Step 2: incorporating the prior information

As in typical Bayesian settings, prior information can and should be incorporated in the ABC analysis. Prior beliefs regarding the parameters and the models (if different models are compared) will be used to modify the information contained in the data to obtain the posterior distributions. These beliefs are incorporated in standard ABC analyses in the simulation step (Step 3), i.e. when the parameters used to simulate each genetic variation data will have values sampled from their prior distributions and the number of simulations performed using each of the different models will be proportional to the prior probability assigned to each model.

Prior distributions should be obviously large enough to include all the values which are considered at least possible and their shape may well vary among parameters. For example, mutation rates are often sampled from gamma distributions, migration rates from exponential distributions, times from uniform or exponential distributions and population size from uniform or log-uniform distribution. These choices may reflect previous knowledge on some parameters (e.g. the gamma distribution usually fit well real mutational data) and/or the need to homogeneously sample the parameter across different orders of magnitude (e.g. migration rates between 10^{-5} and 10^{-1}). Of course, if the value of a parameter is known with relatively high precision [e.g. the starting time of an invasion (Pascual *et al.* 2007)], the parameter should be fixed in the simulations. When different scenarios are compared, they are usually considered with the same prior probability.

Sometimes, prior distributions are slightly modified if 'first shot' simulations produce data sets very different from the observed data. This strategy can be necessary in some circumstances, it can be regarded (Gelman 2008) as a test of prior beliefs when combined with appropriate quality controls (see steps 7 and 9), but it should be honestly and carefully adopted. There is clearly a potential difficulty in using the data twice, both for estimation and to 'refine' the priors and the resulting posterior distributions will not be 'true' Bayesian combinations of prior beliefs and likelihoods.

There are clear computation and logical advantages in using prior distributions and the Bayesian approach compared for example, to maximum likelihood methods, even when the prior knowledge is very limited and consequently flat and wide prior distributions are used (Huelsenbeck *et al.* 2001; Holder & Lewis 2003; Beaumont & Rannala 2004). However, we believe that more efforts should be dedicated to identifying information to incorporate with confidence in the prior distributions, using previous genetic or non-genetic studies. These efforts can be facilitated by the hyperprior approach whereby at least the hyperprior distributions can be narrowed. For multilocus microsatellite data for example, the mean and the variance (the hyper-parameters) of the mutation rates are reasonably well known, whereas the single-locus rates are not. Incorporating robust prior beliefs will produce more accurate and precise estimations and it will also facilitate the interpretation of the results. When prior definitions are based on vague information, the effects of errors in prior beliefs can, and should, be efficiently investigated with a sensitivity analysis within the ABC framework (e.g. Pritchard *et al.* 1999; Estoup *et al.* 2001; Hickerson *et al.* 2006; Verdu *et al.* 2009; Guillemaud *et al.* 2010).

Step 3: choosing the summary statistics

The whole ABC machinery is based on the comparison between observed and simulated data sets and this comparison is made after reducing data sets to summary statistics, SuSt. Unfortunately, there is still no general rule as to which and how many SuSt should be used, although the importance of this step was already recognized since the formal introduction of ABC (Beaumont *et al.* 2002; Marjoram *et al.* 2003). The selected SuSt should be able to capture the relevant features of the data. Ideally, SuSt should be *sufficient*, i.e. the posterior probability of a parameter given these SuSt

should be the same as its posterior distribution given the complete data set (Marjoram & Tavare 2006). In practice, a SuSt should not be included in this set if it does not provide any additional information about the data useful for the estimation process. Easy to say, but very difficult to realize. The sufficiency of a set of SuSt is strictly dependent on the model, parameters and data, meaning that some preliminary analysis is required.

A single or few SuSt are almost always a very crude representation of the data and likely produce biases in ABC analyses (Marjoram *et al.* 2003). On the other hand, too many SuSt (especially those providing little information regarding the parameter being estimated) introduce stochastic noise, reducing the fraction of retained simulations and increasing the errors both when the distance between observed and simulated data sets is estimated and during the regression step (Beaumont *et al.* 2002). More than 100 SuSt were used by Rosenblum *et al.* (2007) to reconstruct the historical demography of a lizard colonization process, but the usual numbers of SuSt in published empirical studies range between 5 and 20.

When several loci are typed, SuSt are usually means and variances of single locus statistics (e.g. Ross-Ibarra *et al.* 2009) or indices correlated to the shape of the distribution of phenotypes (e.g. AFLP data, see Foll *et al.* 2008) or allele (e.g. SNP data) frequencies. At least three methods, not yet implemented for practical use, have been suggested to identify the best set of SuSt. Hamilton *et al.* (2005) used the determination coefficients between each SuSt and each parameter, estimated from a set of preliminary simulations, to weight differentially the SuSt. The distance between observed and simulated data sets is thus computed separately for each parameter. This is not the same as selecting a subset of SuSt, but is a criteria to avoid this selection and to almost exclude by weighting some SuSt from the estimation process. Joyce & Marjoram (2008) have introduced a 'sufficiency' score to be assigned to each SuSt in a sort of preliminary experiment. The whole ABC estimation step is performed several times adding and removing different SuSt and retaining only those that significantly modify the posterior distribution of the parameter of interest. Wegman *et al.* (2009) suggest extracting a limited number of orthogonal components, appropriate to explain the parameters variation, from a large number of SuSt. These new variables, estimated by a partial least square regression approach with coefficients estimated on the basis of a set of preliminary simulations, are then used as SuSt. So far, only modest advantages of these approaches have been demonstrated. Considering the actual state of the art, we recommend a selection of the SuSt known to be informative about the parameters of interest, an appropriate number of simulations (see below) and, in particular, some preliminary tests showing that the selected SuSt can be used to reasonably recover models and parameters in data sets simulated under scenarios relevant for the addressed question (see e.g. Becquet & Przeworski 2007; Rosenblum *et al.* 2007; Neuenschwander *et al.* 2008).

Step 4: simulating the model(s)

A large number of data sets should be simulated under the model(s) defined at Step 1, with each simulation using a different set of parameter values sampled from the corresponding prior distribution. Simulation is the time-consuming step, but an important advantage of standard ABC (but not of ABC coupled with MCMC or importance sampling) is that the data sets generated by simulation can be used for estimation or model selection on many different data sets. The simulated data sets, which are commonly reduced to the values of the chosen SuSt due to disk space limitations, are stored in the *reference table*. The same reference table can then be used for inference on the real data sets but also, for example, on pseudo-observed data sets, or *pods*. Pods are specific data sets generated with known parameter values by simulation and are very useful for investigating the bias/accuracy of the analysis (see steps 7 and 9). The reference table is therefore very valuable, both because it usually takes lot of computing time to generate it and because it will be recycled several times. It seems thus a good idea to select accurately the software for the simulations most appropriate for the scenario and genetic markers of interest, to avoid hurried decisions about the prior distributions and not to economize on the number of simulations.

In principle, both backward coalescent and forward classical simulations of the genetic data can be used for ABC. In practice, only the former seem to have, today, the required time-efficiency. Forward genetic simulations have the advantage to substantially simplify the implementation of natural selection and for this reason they may spread for specific ABC implementations (e.g. Itan *et al.* 2009). We expect that efficient forward genetic simulator (Chadeau-Hyam *et al.* 2008; Hernandez 2008; Carvajal-Rodriguez 2010) coupled with ABC will be used in the near future to analyse complex scenario involving both selective and demographic processes.

The available coalescent simulation programs, reported in Table 1 with their main characteristics, can be classified in two major groups: ABC integrated and ABC independent simulators. ABC integrated simulators are assembled within a larger package designed to perform all the ABC analyses. These user-friendly pack-

Table 1 Features of the main online backward coalescent simulators available for ABC analysis. Other software are developed for specific purposes and available upon request to the authors (see e.g. Przeworski 2003)

		Demographic model									
Name	Type of markers	One/many populations	Population divergence	Migration	Change in population size	Recombination	Selection	Serial sampling[1]	Consider explicitly spatial and environmental heterogeneity	ABC integrated	Reference
MS	DNA sequence	Many	Yes	Yes	Yes	Yes	No[2]	No	No	No	Hudson (2002)
Simcoal2	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	Yes	No	No	No	No	Laval & Excoffier (2004)
Selsim	STR, DNA sequence	One	No	No	No	Yes	Yes	No	No	No	Spencer & Coop (2004)
SPLATCHE	RFLP, STR, DNA sequence[3]	Many	Yes	Yes	Yes	No[3]	No	No	Yes	No	Currat *et al.* (2004)
Bayesian Serial Simcoal	RFLP, STR, DNA sequence	Many	Yes	Yes	Yes	No	No	Yes	No	No[4]	Anderson *et al.* (2005)
AQUASPLATCHE	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	No	No	No	Yes	No	Neuenschwander (2006)
msBayes	DNA sequence	Many[5]	Yes	Yes	Yes	Yes	No	No	No	Yes	Hickerson *et al.* (2007)
DIY ABC	STR, DNA sequence	Many	Yes	No	Yes	No	No	Yes	No	Yes	Cornuet *et al.* (2008)[6]
ONeSAMP	STR	One	No	No	No	No	No	No	No	Yes	Tallmon *et al.* (2008)
PopABC	STR, DNA sequence	Many	Yes	Yes	No	Yes	No	No	No	Yes	Lopes *et al.* (2009)
ABCtoolbox[7]	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes/No[7]	Wegmann *et al.* (2010)

[1]Samples with different ages can be simulated (relevant if ancient DNA data or time series are available).
[2]MS has been modified by Jensen *et al.* (2008) to include selection.
[3]The new version of SPLATCHE, including recombination and SNP markers, will be available soon (L. Excoffier, pers. comm.).
[4]Data can be by simulated sampling parameter values from prior distributions.
[5]Only a vicariance model can be simulated.
[6]The new version of DIYABC is available online.
[7]ABCtoolbox is a collection of independent command-line programs which facilitate the development of a pipeline to estimate model parameters and compare models; several external simulation programs can be pipelined.

ages have obvious advantages since the complete ABC analysis can be accomplished using a single program. Users considering these packages should, however, realize that very complex models with specific prior distributions of the parameters, as well as some kind of genetic markers, may not be simulated. Similarly, other steps of the analysis are constrained to specific functions, whereas the ABC is by nature almost like bricolage, requiring frequent small adjustments (including suggestions coming from new studies) and specific tests for different data-sets. ABC non-integrated simulators are independent programs simulating genetic variation patterns. The best choice is probably to look initially at the general ABC packages with integrated simulators (see Table 1), figure out if the models and markers of interest can be simulated, and, if not, find out if the authors will release an updated version soon (the field is moving fast). However, if some experience with programming and script development is available (e.g. in R, Phyton, or C++), choose an ABC independent simulator. MS (Hudson 2002) and Simcoal2.0 (Laval & Excoffier 2004) are widely used ABC independent simulators. For models which consider explicitly spatial or environmental heterogeneity, i.e. where large numbers of demes and their migration/colonization relationships through time and space are assumed, Splatche (Currat *et al.* 2004) or Aquasplatche (Neuenschwander 2006) are more appropriate. Serial SimCoal (Anderson *et al.* 2005) can be used if ancient DNA data are available. All these ABC independent simulators are very flexible and allow access to the code, but need of course to be 'pipelined' within all the other steps of the ABC analysis. The recent introduction of a series of programs within a single ABC tool box (Wegmann *et al.* 2010) will likely alleviate this problem in the future.

Finally, we have to address another question with no general answer: how many simulations? Empirical studies seem to converge towards the order of magnitude of 10^6. Clearly, the complexity of the model and the dimensions of the parameter affect the number of simulations necessary to explore them. Our view is that some preliminary simulations, for example testing the convergence by comparing the results obtained in a few independent analyses with 10^4–10^5 simulations, can be very useful. In some cases, for example in the relatively simple scenarios analysed by Guillemaud *et al.* (2010) even 10^4 simulations appear sufficient to reach accuracy in model selection. However, if alternative versions of ABC (for example ABC-MCMC) are not considered, brute power rather than style is the main feature of ABC analyses. We suggest therefore using large CPU clusters (now relatively cheap and commonly available in computing departments) and performing several millions of simulations in the final analysis for both model selection and parameter estimates. The current development of specific implementations of genetic analyses using graphics processing units (GPU) (e.g. Suchard & Rambaut 2009) will possibly reduce the need for large clusters soon.

Step 5: filtering the simulations

Simulations are retained when a multivariate distance between observed and simulated SuSt is below a certain distance threshold. In general, a simple Euclidean distance is computed on normalized SuSt and the threshold is defined such that a small fraction of the simulations (0.1–3%), corresponding to the smallest distances, are retained for the estimation step. A conditional threshold (e.g. Thornton & Andolfatto 2006; Putnam *et al.* 2007) can also be devised, implying that simulations (see step 3) are repeated until a certain number (in the order of 10^3–10^4) of accepted simulations is reached. The distance threshold can be different in model selection (step 6) and parameter estimation (step 8). As underlined by Guillemaud *et al.* (2010), the choice of the threshold should be always validated. Pods should be simulated under a model (or models) relevant for the question addressed and the threshold producing reasonably stable and accurate reconstruction of the known scenarios should be adopted. In any case, it is a good idea always to check the effect of using different threshold values on the distribution of Euclidean distances, on the comparison between observed SuSt (separately or combined for example using PCA) and the corresponding simulated distributions and, obviously, on the estimated posterior distributions of the parameters and the models.

Step 6: model selection (if different models are compared)

In step 6 of ABC, some results are obtained at last. Comparing models, which actually means comparing alternative hypotheses about a process, is the key to the work of scientists. The ABC framework allows the computation of the relative weight, i.e. the posterior probability, of different hypotheses (i.e. different models). Probably for the first time, genetic variation data can be used not only to reconstruct a plausible historical or demographic scenario, often combining many different analyses and tests, but also to assign a quantitative 'belief score' to each of many alternative and possibly complex scenarios. ABC *should* also favour a reduction in the length of manuscripts, since elaborated arguments supporting or opposing each hypothesis can be summarized by a corresponding set of meaningful probability scores, the sum of which is always equal to

one. The fear is that this reduction will be compensated by extensive, technical and somewhat boring description of ABC options and validation, but general readers will be able to skip these sections and easily comprehend the results and main conclusions.

All the models are generally simulated the same number of times. This is equivalent to giving the same prior probability to each model under comparison and zero probability to any other model. Clearly, errors in the latter assumption may produce incorrect conclusions regarding the models support (see e.g. Templeton 2009), but the ABC framework allows for the evaluation of the effects of excluding some models in the specific situation under investigation (Guillemaud *et al.* 2010). In the final set of retained simulations, the data sets produced by the more probable models will be over-represented and the data sets produced by the less probable models will be under-represented or even absent. Intuitively, the probability of a model is proportional to the relative frequency of the data sets it produces that are among the retained simulations (Weiss & von Haeseler 1998; Pritchard *et al.* 1999). This frequency is actually the direct estimator of the posterior probability of a model, but this estimator is rarely accurate in complex scenarios when, inevitably, the retained simulations are either too few or also contain data sets not closely matching the observed data. Recently, Leuenberger & Wegmann (2010) proposed the use of a parametric General Linear Model to adjust the model frequencies in the retained simulations. However, the most reliable and tested method, also available in ABC packages such as DIYABC (Cornuet *et al.* 2008), is still the adjustment based on the weighted multinomial logistic regression introduced by Beaumont (2008). The coefficients for the regression between a model indicator (response) variable and the simulated SuSt (the explanatory variables) can be estimated, allowing the estimation of the posterior probability for each model at the intercept condition where observed and simulated SuSt coincide. CIs of the probabilities can be computed as suggested by Cornuet *et al.* (2008).

The posterior probability of each model is of course an intuitive score of our belief in that model. An additional index, comparable with a standard table of reference values where the evidence is assigned to categories from 'not worth more than a bare mention' to 'decisive', is the Bayes factor. The Bayes factor can be easily computed in an ABC analysis, being the ratio between the posterior probabilities estimated in any pair of models, divided by the ratio of their prior probabilities. The latter ratio is of course equal to one if all models have the same prior probability. This index is a summary of the evidence provided by the data if favour of a model as opposed to another and it can be interpreted as a measure of the relative success of the models at predicting the data (Kass & Raftery 1995). The Bayes factor is also the ratio of the posterior odds to the prior odds, meaning that it actually measures the change of relative probabilities of the various scenarios tested in the ABC analysis due to the knowledge obtained from the genetic data.

Step 7: quality control in model selection

The ABC framework can be used to investigate the robustness of model selection and parameter estimation with relatively little additional effort (e.g. Fagundes *et al.* 2007; Guillemaud *et al.* 2010). Data sets simulated under specific scenarios with known parameter values are tested against the same reference table (the large number of simulated data sets, see step 4) used in the analysis of the real data set.

Some hundreds of pseudo-observed data sets (the pods, see step 3) are generated using each of the scenarios considered in the model selection analysis. Obviously, other scenarios can be analysed to investigate the effects of incomplete model specification on the inference (see step 6). The values of the parameters used for generating pods are generally restricted to the best estimates obtained from the analysis of the real data, but they can be other values of interest. Pods generated with fixed parameters will provide information about the quality of the estimated model probabilities which is restricted to a specific parameter set. The ability of ABC to identify the correct model in a larger space of parameter values can be analysed by generating pods using parameter values sampled from, for example, the prior distribution (Fagundes *et al.* 2007; Cornuet *et al.* 2008; Verdu *et al.* 2009) or the posterior distributions estimated from the real data set.

Even if the definitions here are not rigorous, type I and type II errors can be estimated for each scenario, using, in turn, each scenario as the null or alternative hypothesis. In practice, the type I error for, say, scenario A is estimated as the fraction of pods generated under scenario A that support other scenarios, whereas the type II error for scenario A is estimated as the fraction of pods generated under all the other scenarios that support scenario A. A single pod is considered to be supporting a scenario simply if the posterior probability of this scenario is the largest. So, these are not really type I and II errors in the classical frequestist framework, whereby the null hypothesis is never accepted and is rejected only if the data are manifestly incompatible with it. These estimated errors can be very useful when small, but otherwise their joint interpretation may not be straightforward. Some additional insight into the accuracy and power of the analysis can be obtained by

computing mean and standard deviation of the posterior probability of each model using the probabilities estimated in the pods or the frequency of pods where the CI of the model with the highest probability does not overlap with the CI of the next supported model (see Guillemaud *et al.* 2010). Pods could be also used to estimate the distribution of the Bayes factor for each simulated scenario also and thus to better interpret the Bayes factor computed from the real data set.

Step 8: parameters estimation

For the single model considered in the analysis or for the most supported model if different models are considered, the posterior distributions of the parameters can be reconstructed by ABC. In many cases it is a good idea to start looking at the estimated distribution of the hyper-parameters or the composite parameters, the latter obtained by combining, in each simulation, the parameters which are difficult to estimate separately (for example, the population-mutation parameter $N\mu$ which combines the population size N and the mutation rate μ).

As already explained, retained simulations have data sets closer (but not identical) to the real data than do non-retained simulations. Therefore, using the parameter values of the retained simulations as a sample from their posterior probability distribution (the direct approach), still maintains an undesirable component of the prior. If all simulations are retained, i.e. with a threshold of tolerance equal to infinity, the prior will be recovered. At the other extreme, when the threshold is proximate to 0, the direct approach works well, but huge numbers of simulations are needed to obtain a reasonable sample size from the posterior. We can imagine that in the near future, especially if different groups will share their reference tables, billions of simulations will be available for the direct method. In the meantime, the most commonly used method to adjust the imperfect retained simulations is the local linear weighted regression introduced by Beaumont *et al.* (2002). The coefficients of a linear regression between each parameter and the vector of the chosen SuSt are estimated from the retained simulations (the local aspect) assigning to each point a weight based on a function increasing as the distance between the observed and simulated data sets decreases (the weighting aspect). The regression slope is then used to adjust each parameter value from the retained simulations towards the value expected in correspondence with the observed SuSt. The intercept corresponds to the posterior mean estimate of the parameter. This approach, which can be applied to all the parameters simultaneously, assumes local linearity between parameters and SuSt (see Blum & Francois 2009 for an extension to non-linear regression models), additivity and a multivariate normal distribution. However, its use in the last 8 years after the original development (Beaumont *et al.* 2002) suggest that small violations of these assumptions only marginally affect the results. The accuracy of the posterior distributions, when evaluated under simple scenarios which allows also the use of full-likelihood methods, is drastically increased by the regression step compared to the direct approach (Beaumont *et al.* 2002; Leuenberger & Wegmann 2010). The commonly-used weighting function is the Epanechnikov kernel, but the effects on the final estimates of applying other weighting schemes are probably limited. Parameters are usually transformed before the regression step (and then back transformed after it) by a log (e.g. Estoup *et al.* 2004; Hamilton *et al.* 2005; Crestanello *et al.* 2009), logtan (e.g. Kayser *et al.* 2008; Ross-Ibarra *et al.* 2008) or logit function (e.g. Cornuet *et al.* 2008). Logtan and logit functions avoid adjustments outside the prior distribution.

The general linear model (GLM) approach recently proposed by Leuenberger & Wegmann (2010) and implemented in ABCtoolbox (Wegmann *et al.* 2010) can be used as an alternative method to estimate the posterior distributions from the retained simulations. Additional testing is however necessary to identify the best adjustment procedure under different conditions, since GLM have both advantages (it considers the correlation among SuSt and never produces posterior distributions outside the priors) and disadvantages (it assumes normal distributions for the SuSt and is computationally more intensive) compared to the earlier approach.

Of course, when a sample from the posterior distribution is available, point estimators and a relative measure of accuracy are needed. Usually, a smoothed-posterior density is fitted to the sample of adjusted parameter values using specific methods (e.g. local-likelihood) and after specifying a bandwidth. This fitting step is embedded in the DIYABC package (Cornuet *et al.* 2008), but we suggest analysing the rough-frequency distribution of adjusted parameters to identify possible distortions introduced by the fitting algorithm.

The point estimators usually computed from the posterior densities are the mean, the mode, the median and the intercept estimated in the regression step. No consensus has been reached about the point estimator with the smallest bias and variance and, as usual, the analysis of simulated data sets relevant for the scenario of interest can be useful. In general, however, the differences between point estimators are quite small if compared with the width of their posterior distribution and their choice is therefore almost irrelevant. Most importantly, the posterior density can be used to compute the confidence of the estimates. Typical measures are the

SD and the credible interval, the latter being the Bayesian equivalent of the frequentist CI. A commonly used credible interval is the X% highest posterior density or HPD. The X% HPD interval is the interval which includes the X% of the parameter values and within which the density is never lower than the density outside it. Typically, 90 or 95 HPD limits are reported in ABC analyses (e.g., Fagundes *et al.* 2007; Ludwig *et al.* 2008; Ray *et al.* 2010).

Step 9: quality control of the model and the parameters estimation

The quality of the parameter estimates can be initially evaluated by the proportion of parameter variance that is explained by SuSt (see e.g. Fagundes *et al.* 2007; Neuenschwander *et al.* 2008). This is the classical coefficient of determination. If only a small fraction of parameter variation in the reference table can be explained by variation in SuSt, it is hard to imagine that the parameter will be accurately estimated for the model(s) under consideration and the number of individuals and markers typed. It is possible that different SuSt might improve the estimation (and this hypothesis can be tested), but it is also possible that the parameter cannot be estimated for that model/data package even if the full likelihood could be computed. The coefficient of determination should be taken with caution. Even a small fraction of the explained variation can be sufficient for reasonably precise estimates given enough data. In analogy, population assignment of single individuals can be quite accurate even when a small fraction of the variation is attributed to between-population differences (e.g. Latch *et al.* 2006; Colonna *et al.* 2009).

A more important evaluation of the quality of parameter estimates under the specific scenario (or scenarios) under investigation can be performed within the ABC framework in exactly the same way we described for model selection: generating pods, i.e. simulating datasets using known parameter values or parameter distributions and analysing them (e.g. Excoffier *et al.* 2005; Jensen *et al.* 2008). The best estimates of the parameters obtained in step 8 (or their posterior distributions) are of course interesting candidates for generating pods in this analysis. For each pods, parameters are estimated using the same reference table and the same procedure applied to the real data and are then compared to the true known values used to generate them. In fact, after the analysis of, say, 1000 pods, 1000 posterior densities will be available for each parameter. From each of these distributions, a point estimator (e.g. the mode) and a credible interval (e.g. the 90% HPD) can be computed and several measures of the estimator quality can be estimated (see e.g. Cornuet *et al.* 2008) by simply comparing these 1000 point estimators and credible intervals with the true value of the parameter (i.e. the value used in the simulations). The relative bias, the coefficient of variation and the 90% or the 50% coverage (which are the fraction of 90% or 50% HPD intervals in the 1000 simulations which include the true value), are usually informative to ascertain the quality of the estimates. The analysis and interpretation of other highly-correlated measures is likely useless and confounding. Some caution is also needed in general for the interpretation of these performance measures. For example, a relative bias of 1 (100%) when estimating a true population size of 1000, meaning that on the average the estimated value will lay at a distance of 1000 individuals from the true value, would appear enormous. But if the prior knowledge on this parameter was entirely missing and a uniform prior distribution ranging between 100 and 100 000 was defined, such bias should be considered small. If possible, it is always very instructive to estimate at least some parameters from the pods using other non-ABC approaches and then compare the quality measures across methods (Guillemaud *et al.* 2010). A large bias or variance using ABC can become acceptable in comparison with the performance of other methods.

A third way to investigate the quality of the ABC results is to compare the SuSt observed in the real data with the posterior distribution of SuSt (e.g. Pascual *et al.* 2007; Ingvarsson 2008). The posterior distribution of SuSt is the SuSt distribution computed from pods generated by simulation with parameters values sampled from their estimated posterior distribution (Gelman *et al.* 2004). The rationale of this comparison, which is testing the goodness-of-fit of the combination 'scenario + posterior distributions of the parameters' to the data, is simple: if the estimated parameters under a specific model have anything to do with what happened in the history of the real samples, then histories simulated using these values should produce pods similar to the data. If this is not the case, either the parameter estimation is bad and/or the model is wrong. Using a simple graphical inspection, the goodness-of-fit should be considered high if the distance between observed SuSt and the SuSt in their posterior distribution is low. A principal component analysis of the SuSt in the real data set, the SuSt from their posterior distribution and also the SuSt in the reference table can provide additional insights on the quality of the estimation (Guillemaud *et al.* 2010; A. Estoup, pers. comm.). Also, the performance of the estimation can be quantified by computing bias and variance, relative to the observed SuSt, of the posterior distributions of SuSt (see e.g. Neuenschwander *et al.* 2008).

The comparison between observed SuSt and their posterior distributions is also the basis for a posterior

predictive test, which is the Bayesian analogous of the parametric bootstrap under the frequentist framework (Gelman *et al.* 2004). The goodness-of-fit of the inferred combination 'scenario + the posterior distribution of the parameters' and the observed data is quantitatively measured by the posterior predictive *P* value. This bayesian *P* value corresponds to the probability that data replicated using the estimated posterior distributions of the parameters are more extreme that the observed data (Gelman *et al.* 2004). It can be specific for each SuSt (e.g. Thornton & Andolfatto 2006) or appropriately combined across SuSt (e.g. Ghirotto *et al.* 2010) and it can be also be viewed as the probability of observing a less good fit between the model and the data.

A posterior predictive test is a good practice in any Bayesian model-based analysis, since it is the most straightforward way to understand if the estimated parameters are at least meaningful. In general, however, its power to reject the null hypothesis under reasonable population genetic condition and particularly when few loci are analysed, is limited. Nonetheless, this test can be useful to identify deviant SuSt with significant *P* values, possibly related to specific poorly estimated parameters or erroneous aspects of the demographic model. The use of posterior predictive tests in ABC is therefore recommended (e.g. Becquet & Przeworski 2007; Ingvarsson 2008; Neuenschwander *et al.* 2008; Ghirotto *et al.* 2010).

Applications

The number of published applications of ABC to genetic variation data increased rapidly following the formal definition of the methodology in 2002 (Beaumont *et al.* 2002), doubling for example between 2007 and 2008. A bibliographic Endnote list of 107 papers on ABC, with about two-thirds of them presenting applications to real data, is provided as 'Supporting information'.

Approximate Bayesian computation has been applied to very different types of organisms, from bacteria (e.g. Luciani *et al.* 2009; Wilson *et al.* 2009) to plants (e.g. Francois *et al.* 2008; Ross-Ibarra *et al.* 2009) and animals (e.g. Voje *et al.* 2009; Lopes & Boessenkool 2010). Microsatellite markers are the most commonly used source of genetic information (43% of the studies), followed by nuclear and mitochondrial DNA sequences (each of them analysed in about 30% of the studies). DNA data from ancient samples is included in about 10% of the studies. The number of loci varies widely among papers, but the median value for STR and nuclear sequences is 9 and 19, respectively.

After surveying the many options available when running an ABC analysis, we outline in the 'Supporting information' the main trends. In general, we estimated that if all the steps discussed in the previous section were applied, about 60% of the published ABC applications would have significantly improved their robustness. As positive examples of studies in the field of molecular ecology where, in our opinion, the ABC framework was properly used to estimate parameters, to compare models and to evaluate the quality of the model settings and the results, we would like to mention Neuenschwander *et al.* (2008) and Guillemaud *et al.* (2010). Neuenschwander *et al.* (2008) reconstructed the dynamic of the post-glacial colonization of a river basin in Switzerland by the European bullhead (*Cottus gobio*). Guillemaud *et al.* (2010), after extensively investigating the capabilities of ABC in reconstructing different aspects of an invasion process, applied this method to investigate alternative scenarios for the introduction to Europe of the North American pest of corn *Diabrotica virgifera*.

Using a subset of 14 relatively homogeneous studies and 152 parameter estimates, we also compared prior and posterior distributions to obtain some general indications about the fraction of uncertainty about a species that was reduced by ABC using genetic information. The difference in width and location between prior and posterior distributions was not clearly related to any general features of the data sets, the model or the ABC setting (such as the number of loci, the number of sampled individuals, the number of parameters to be estimated and the number of SuSt). If applied to a single study, this result would appear counter intuitive, since increasing for example the number of markers should narrow the credible intervals (Excoffier *et al.* 2005). The power of our analysis is clearly limited, but it is also possible that large differences in the informativeness of the data sets and in the complexity of the scenarios across studies blurred the expected pattern. At any rate, our analysis of more than 150 posterior distributions seems to confirm the reasonable idea that guidelines regarding the number of individuals and markers to analyse, the maximum allowed complexity of the model and the number of SuSt sufficient to summarize the data, cannot be easily identified. Such guidelines can be very specific only for the process and species of interest, meaning that the set of preliminary simulations described throughout this paper can be very useful to plan the sampling, the typing and the final ABC setting. Additional results and details of this analysis are provided as 'Supporting information', Table S1 and Fig. S1.

Conclusions

Approximate Bayesian computation has a short history and very likely a long future. Molecular variation data

are useful to reconstruct past events, estimate biological parameters and compare alternative scenarios. The ABC approach has the potential to become a standard approach in molecular ecology, as well as in other fields (Lopes & Beaumont 2010; Lopes & Boessenkool 2010). It allows, for the first time, the efficient exploitation of the enormous progresses in genetic typing and computing speed to investigate very complex population models including both natural and human-induced processes.

Throughout this paper we have summarized the theoretical and practical aspects of this methodology, which should not be considered as a statistical analysis *per se* but rather as a statistical framework. We also briefly analysed its behaviour reviewing the published applications. Overall, we tried to stimulate the general reader to consider ABC as a possible instrument for analysing their data and also for planning sampling and typing strategies. Many pros and cons, together with some practical suggestions, were given. We schematically recall and integrate them in this final section.

Complex and specific models can be analysed

The likelihood of models and parameters does not need to be theoretically derived. We believe that, to a reasonable extent, initial investigations under the ABC framework should always take advantage of this quality and challenge the data using very realistic models. Our analysis of empirical studies suggests that even a few markers can be useful to substantially increase the knowledge about the process of interest. This is possible in many cases by limiting the inference to well-estimated composite and hyper-parameters. Importantly, ABC provides the instruments to understand if the set data + models can be used or not to reconstruct the most likely scenario and if the estimate can progress from composite to single parameters.

Quality of estimates and model selection can be measured

After becoming familiar with the ABC framework, quality controls and power analyses can be performed with rather limited additional effort. The simulation results, stored in a single reference table, allow the analysis of the real data set as well as many other simulated data sets of interest and the feasibility of an accurate estimation or a model selection can be analysed. It is recommended that large reference tables stored by different research groups become accessible, possibly promoting a shared repository, since some preliminary analysis before the sampling and typing using simulated data sets or real data sets with properties (sample sizes, number and type of loci, etc.) matching as closely as possible the data sets in the reference table, could be very useful. Interestingly, all these analyses in specific ABC settings for different species and questions will also help in understanding the general properties of the method. Unfortunately, quality control is more difficult under the ABC-MCMC and related serial approaches, since the dependency among simulated data sets prevents the compilation of a reference table.

Difficulties in performing an ABC analysis are decreasing

The conceptual scheme of ABC is quite simple and modular, implying that end users who do not find enough flexibility in complete ABC packages [e.g. DIY-ABC, Cornuet *et al.* (2008)] can develop specific implementations using the appropriate simulator (e.g. MS) pipelined with relatively simple algebraic or statistical computations. The post-simulation analyses are already coded in available and easily modifiable R scripts or C/C++ programs. A great collection of command-line modules useful for an ABC analysis, which includes also samplers based on MCMC (Marjoram *et al.* 2003) and PMC (Beaumont *et al.* 2009), is also available in the ABCtoolbox (Wegmann *et al.* 2010). Important ABC implementations for estimating for example selection coefficients or individual-based parameters will likely imply the development of more efficient simulators (see e.g. Przeworski 2003; Jensen *et al.* 2008; Leblois *et al.* 2009).

Probabilities are approximated

More studies are needed to better evaluate the degree of approximation of ABC estimates compared to full-data likelihood methods. These studies, however, are restricted to scenarios where explicit likelihoods functions are tractable. Fortunately, the ABC approach has the potential to be used, case by case and if properly implemented, as a self-evaluator of its performances under known simulated scenarios. This potential, clearly related to the fact that an ABC application is actually a hybrid between a simulation and a data-analysis study, can be used to understand general questions regarding, for example, the ability of ABC to select the true among many simulated models even when the parameters of the model are poorly estimated.

Time and patience are required

It is very important to realize that performing an ABC study is not at all like using other methods for the analysis of genetic variation data. Even if questions of interest can be addressed using a single complete package such

as DIYABC (Cornuet *et al.* 2008), adequate amount of time should be dedicated to initially define the model(s) and then decide among the many alternative options. Crucial steps such as deciding on the number of simulations, the SuSt and the acceptance threshold cannot be based on general rules. The effects of these choices and the performances of the estimates should be evaluated and tested in each study. Fortunately, even if planning a rigorous ABC analysis in all its steps, as well as modifying the initial plan when needed, requires time and patience, computer speed (for example exploiting hundreds of CPUs and possibly, in the future, GPUs) and data storage possibilities (terabytes are very cheap today) are not a limiting factor in many studies. Elegant methods to efficiently reduce the number of relevant simulations needed for the estimation process, such as ABC coupled with MCMC, are under development and testing, but their standardization and spread to non-experts might not be rapid.

Acknowledgements

We thank Arnaud Estoup and two anonymous reviewers for useful and detailed comments and suggestions. This research was supported by the Italian Ministry for Research and Education and by the University of Ferrara, Italy.

References

Anderson CN, Ramakrishnan U, Chan YL, Hadly EA (2005) Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics*, **21**, 1733–1734.

Beaumont MA (2008) Joint determination of topology, divergence time and immigration in population trees. In: *Simulation, Genetics and Human Prehistory* (eds Matsamura S, Forster P, Rrenfrew C), pp. 135–154. McDonald Institute for Archaeological Research, Cambridge, UK.

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nature Review Genetics*, **5**, 251–261.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Beaumont MA, Cornuet J-M, Marin JM, Robert CP (2009) Adaptivity for ABC algorithms: the ABC-PMC scheme. *Biometrika*, **96**, 983–990.

Beaumont MA, Nielsen R, Robert C *et al.* (2010) In defence of model-based inference in phylogeography. *Molecular Ecology*, **19**, 436–446.

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**, 1505–1519.

Blum MGB, Francois O (2009) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, **20**, 63–73.

Bortot P, Coles SG, Sisson SA (2007) Inference for stereological extremes. *Journal of the American Statistical Association*, **102**, 84–92.

Cappè O, Guillin A, Marin JM, Robert C (2004) Population Monte Carlo. *Journal of Computing Graphics and Statistics*, **13**, 907–929.

Carvajal-Rodriguez A (2010) Simulation of genes and genomes forward in time. *Current Genomics*, **11**, 58–61.

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF *et al.* (2008) Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, **9**, 364.

Colonna V, Nutile T, Ferrucci RR *et al.* (2009) Comparing population structure as inferred from genealogical versus genetic information. *European Journal of Human Genetics*, **17**, 1635–1641.

Cornuet JM, Santos F, Beaumont MA *et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Crestanello B, Pecchioli E, Vernesi C *et al.* (2009) The genetic impact of translocations and habitat fragmentation in chamois (*Rupicapra*) spp. *Journal of Heredity*, **100**, 691–708.

Currat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, **4**, 139–142.

Doucet A, de Freitas JFG, Gordon NJ (2001) *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, NY.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

Estoup A, Clegg SM (2003) Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*. *Molecular Ecology*, **12**, 657–674.

Estoup A, Wilson IJ, Sullivan C, Cornuet JM, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet JM (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution*, **58**, 2021–2036.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Excoffier L, Estoup A, Cornuet JM (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.

Fagundes NJ, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences, USA*, **104**, 17614–17619.

Foll M, Beaumont MA, Gaggiotti O (2008) An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study population structure. *Genetics*, **179**, 927–939.

Francois O, Blum MG, Jakobsson M, Rosenberg NA (2008) Demographic history of european populations of Arabidopsis thaliana. *PLoS Genetics*, **4**, e1000075.

Fu YX, Li WH (1997) Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution*, **14**, 195–199.

Gelman A (2008) *Rejoinder. Bayesian Analysis*, **3**, 467–478.
Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC, London, UK.
Ghirotto S, Mona S, Benazzo A *et al.* (2010) Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. *Molecular Biology and Evolution*, **27**, 875–886.
Guillemaud T, Beaumont MA, Ciosi M, Cornuet JM, Estoup A (2010) Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, **104**, 88–99.
Hamilton G, Currat M, Ray N *et al.* (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics*, **170**, 409–417.
Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.
Hickerson MJ, Stahl EA, Lessios HA (2006) Test for simultaneous divergence using approximate Bayesian computation. *Evolution*, **60**, 2435–2453.
Hickerson MJ, Stahl E, Takebayashi N (2007) msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics*, **8**, 268.
Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature Review Genetics*, **4**, 275–284.
Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.
Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.
Ingvarsson PK (2008) Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics*, **180**, 329–340.
Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The origins of lactase persistence in Europe. *PLoS Computing and Biology*, **5**, e1000491.
Jensen JD, Thornton KR, Andolfatto P (2008) An approximate Bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. *PLoS Genetics*, **4**, e1000198.
Joyce P, Marjoram P (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7**, 26.
Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
Kayser M, Lao O, Saar K *et al.* (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *American Journal of Human Genetics*, **82**, 194–198.
Kimura M, Weiss GH (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561–576.
Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics*, **140**, 1421–1430.
Latch E, Dharmarajan G, Glaubitz J, Rhodes O (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, **7**, 295–302.
Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–2487.
Leblois R, Estoup A, Rousset F (2009) IBDSim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, **9**, 107–109.
Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics*, **184**, 243–252.
Lopes JS, Beaumont MA (2010) ABC: a useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, In Press.
Lopes JS, Boessenkool S (2010) The use of approximate Bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. *Conservation Genetics*, **11**, 421–433.
Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. *Bioinformatics*, **25**, 2747–2749.
Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM (2009) The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences, USA*, **106**, 14711–14715.
Ludwig A, Arndt U, Lippold S *et al.* (2008) Tracing the first steps of American sturgeon pioneers in Europe. *BMC Evolutionary Biology*, **8**, 221.
Marjoram P, Tavare S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Review Genetics*, **7**, 759–770.
Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA*, **100**, 15324–15328.
Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. *Science*, **201**, 786–792.
Neuenschwander S (2006) AQUASPLATCHE: a program to simulate genetic diversity in populations living in linear habitats. *Molecular Ecology Notes*, **6**, 583–585.
Neuenschwander S, Largiader CR, Ray N *et al.* (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Molecular Ecology*, **17**, 757–772.
Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.
Pascual M, Chapuis MP, Mestres F *et al.* (2007) Introduction history of Drosophila subobscura in the New World: a microsatellite-based survey using ABC methods. *Molecular Ecology*, **16**, 3069–3083.
Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.
Przeworski M (2003) Estimating the time since the fixation of a beneficial allele. *Genetics*, **164**, 1667–1676.
Putnam AS, Scriber JM, Andolfatto P (2007) Discordant divergence times among Z-chromosome regions between two ecologically distinct swallowtail butterfly species. *Evolution*, **61**, 912–927.
Ray N, Wegmann D, Fagundes NJ *et al.* (2010) A statistical evaluation of models for the initial settlement of the

American continent emphasizes the importance of gene flow with Asia. *Molecular Biology and Evolution*, **27**, 337–345.
Rosenblum EB, Hickerson MJ, Moritz C (2007) A multilocus perspective on colonization accompanied by selection and gene flow. *Evolution*, **61**, 2971–2985.
Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, **3**, e2411.
Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the genus Zea. *Genetics*, **181**, 1399–1413.
Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA*, **104**, 1760–1765.
Sokal RR, Oden NL, Barker JSF (1987) Spatial structure in *Drosophila buzzatii* populations: simple and directional spatial autocorrelation. *American Naturalist*, **129**, 122–142.
Spencer CC, Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, **20**, 3673–3675.
Suchard MA, Rambaut A (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatics*, **25**, 1370–1376.
Tallmon DA, Koyuk A, Luikart GH, Beaumont MA (2008) ONeSAMP: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources*, **8**, 299–301.
Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.
Templeton AR (1993) The Eve hypotheses: a genetic critique and reanalysis. *American Anthropologist*, **95**, 51–72.
Templeton AR (2009) Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology*, **18**, 319–331.
Templeton AR (2010a) Coalescent-based, maximum likelihood inference in phylogeography. *Molecular Ecology*, **19**, 431.
Templeton AR (2010b) Coherent and incoherent inference in phylogeography and human evolution. *Proceedings of the National Academy of Sciences, USA*, **107**, 6376–6381.
Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. *Genetics*, **172**, 1607–1619.
Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202.
Verdu P, Austerlitz F, Estoup A *et al.* (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, **19**, 312–318.
Voje KL, Hemp C, Flagstad O, Saetre GP, Stenseth NC (2009) Climatic change as an engine for speciation in flightless Orthoptera species inhabiting African mountains. *Molecular Ecology*, **18**, 93–108.
Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.
Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.
Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.
Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics*, **149**, 1539–1546.
Wilson DJ, Gabriel E, Leatherbarrow AJ *et al.* (2009) Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular Biology and Evolution*, **26**, 385–397.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 The ratio of posterior vs. prior distribution range (RR) and the ratio of the largest vs. the smallest location measure in the posterior and the prior distributions (ER) computed from 14 ABC studies and 152 parameter estimates. n: number of parameter estimates subdivided into four groups; Q1 and Q3: first and third quartile. Sample sizes are not the same for RR and ER because the information required to compute them was not homogeneous across studies. See text for additional details

Fig. S1 The relationship between ER and RR when the median value of ER is computed separately within six RR bins. Bars are quartiles, and the numbers indicate the fraction of point belonging to each bin. Codes are as in Table 2.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

BIBLIOGRAFIA

Alves, D., V. Imperatriz-Fonseca, et al. (2011). "Successful maintenance of a stingless bee population despite a severe genetic bottleneck." Conservation Genetics **12**(3): 647-658.

Anderson, C. N., U. Ramakrishnan, et al. (2005). "Serial SimCoal: a population genetics model for data from multiple populations and points in time." Bioinformatics **21**(8): 1733-1734.

Atkinson, Q. D., R. D. Gray, et al. (2008). "mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory." Mol Biol Evol **25**(2): 468-474.

Atkinson, Q. D., R. D. Gray, et al. (2009). "Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa." Proc Biol Sci **276**(1655): 367-373.

Banks, S. C., S. D. Ling, et al. (2010). "Genetic structure of a recent climate change-driven range extension." Mol Ecol **19**(10): 2011-2024.

Beaumont, M. (2008). Joint determination of topology, divergence time, and immigration in population trees. Simulation, Genetics, and Human Prehistory. F. P. Matsumura S, Renfrew C, editors. Cambridge, McDonald Institute for Archaeological Research**:** 135-154.

Beaumont, M. A. (1999). "Detecting population expansion and decline using microsatellites." Genetics **153**(4): 2013-2029.

Beaumont, M. A. (2010). "Approximate Bayesian Computation in Evolution and Ecology." Annual Review of Ecology, Evolution, and Systematics **41**(1): 379-406.

Beaumont, M. A., W. Zhang, et al. (2002). "Approximate Bayesian computation in population genetics." Genetics **162**(4): 2025-2035.

Beerli, P. and J. Felsenstein (2001). "Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach." Proc Natl Acad Sci U S A **98**(8): 4563-4568.

Campos, P. F., E. Willerslev, et al. (2010). "Ancient DNA analyses exclude humans as the driving force behind late Pleistocene musk ox (Ovibos moschatus) population dynamics." Proc Natl Acad Sci U S A **107**(12): 5675-5680.

Carvajal-Rodriguez, A. (2008). "GENOMEPOP: a program to simulate genomes in populations." BMC Bioinformatics **9**: 223.

Chan, Y. L., C. N. Anderson, et al. (2006). "Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA." PLoS Genet **2**(4): e59.

Chikhi, L., V. C. Sousa, et al. (2010). "The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes." Genetics **186**(3): 983-995.

Cockerham, C. C. and B. S. Weir (1984). "Covariances of relatives stemming from a population undergoing mixed self and random mating." Biometrics **40**(1): 157-164.

Crow, J. and M. Kimura (1970). An Introduction to Population Genetics Theory, Blackburn Press.

Csillery, K., M. G. Blum, et al. (2010). "Approximate Bayesian Computation (ABC) in practice." Trends Ecol Evol **25**(7): 410-418.

Doucet, A., N. De Freitas, et al. (2001). Sequential Monte Carlo methods in practice. New York, NY, Springer-Verlag.

Drummond, A. J., G. K. Nicholls, et al. (2002). "Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data." Genetics **161**(3): 1307-1320.

Drummond, A. J. and A. Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees." BMC Evol Biol **7**: 214.

Drummond, A. J., A. Rambaut, et al. (2005). "Bayesian coalescent inference of past population dynamics from molecular sequences." Mol Biol Evol **22**(5): 1185-1192.

Dudek, S. M., A. A. Motsinger, et al. (2006). "Data simulation software for whole-genome association and other studies in human genetics." Pac Symp Biocomput: 499-510.

Eastman, J. T. (2005). "The nature of the diversity of Antarctic fishes." Polar Biology **28**(2): 93-107.

Eastman, J. T. and R. R. Eakin (2000). "An updated species list for notothenioid fish (Perciformes ; Notothenioidei), with comments on Antarctic species." Archive of Fishery and Marine Research **48**(1): 11--20.
Eastman, J. T. and A. R. McCune (2000). "Fishes on the Antarctic continental shelf: evolution of amarine species flock?*." Journal of Fish Biology **57**: 84-102.
Estoup, A., P. Jarne, et al. (2002). "Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis." Mol Ecol **11**(9): 1591-1604.
Excoffier, L., A. Estoup, et al. (2005). "Bayesian analysis of an admixture model with mutations and arbitrarily linked markers." Genetics **169**(3): 1727-1738.
Excoffier, L. and H. E. Lischer (2010). "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows." Mol Ecol Resour **10**(3): 564-567.
Excoffier, L., J. Novembre, et al. (2000). "SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography." J Hered **91**(6): 506-509.
Fagundes, N. J., N. Ray, et al. (2007). "Statistical evaluation of alternative models of human evolution." Proc Natl Acad Sci U S A **104**(45): 17614-17619.
Felsenstein, J. (2006). "Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci?" Mol Biol Evol **23**(3): 691-700.
Ficetola, G. F., A. Bonin, et al. (2008). "Population genetics reveals origin and number of founders in a biological invasion." Mol Ecol **17**(3): 773-782.
Finlay, E. K., C. Gaillard, et al. (2007). "Bayesian inference of population expansions in domestic bovines." Biol Lett **3**(4): 449-452.
Fisher, R. (1930). The Genetic Theory of Natural Selection. Oxford, Clarendon Press.
Francois, O., M. G. Blum, et al. (2008). "Demographic history of european populations of Arabidopsis thaliana." PLoS Genet **4**(5): e1000075.
Fu, Y. X. (1997). "Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection." Genetics **147**(2): 915-925.
Fu, Y. X. and W. H. Li (1997). "Estimating the age of the common ancestor of a sample of DNA sequences." Mol Biol Evol **14**(2): 195-199.
Garza, J. C. and E. G. Williamson (2001). "Detection of reduction in population size using data from microsatellite loci." Mol Ecol **10**(2): 305-318.
Gelman, A., J. B. Carlin, et al. (2003). Bayesian data analysis, Chapman and Hall/CRC.
Ghirotto, S., F. Tassi, et al. (2011). "No evidence of Neandertal admixture in the mitochondrial genomes of early European modern humans and contemporary Europeans." Am J Phys Anthropol **146**(2): 242-252.
Godoy, J. A., J. J. Negro, et al. (2004). "Phylogeography, genetic structure and diversity in the endangered bearded vulture (Gypaetus barbatus, L) as revealed by mitochondrial DNA." Mol Ecol **13**(2): 371-390.
Guillemaud, T., M. A. Beaumont, et al. (2010). "Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data." Heredity (Edinb) **104**(1): 88-99.
Hamilton, G., M. Currat, et al. (2005). "Bayesian estimation of recent migration rates after a spatial expansion." Genetics **170**(1): 409-417.
Hansen, M. M. (2002). "Estimating the long-term effects of stocking domesticated trout into wild brown trout (Salmo trutta) populations: an approach using microsatellite DNA analysis of historical and contemporary samples." Mol Ecol **11**(6): 1003-1015.
Hansen, M. M., D. J. Fraser, et al. (2008). "Reproductive isolation, evolutionary distinctiveness and setting conservation priorities: the case of European lake whitefish and the endangered North Sea houting (Coregonus spp.)." BMC Evol Biol **8**: 137.
Hasegawa, M., H. Kishino, et al. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." Journal of Molecular Evolution **22**(2): 160-174.
Hein, J., M. Schierup, et al. (2005). Gene Genealogies, Variation and Evolution : A Primer in Coalescent Theory, {Oxford University Press, USA}.

Heled, J. and A. J. Drummond (2008). "Bayesian inference of population size history from multiple loci." BMC Evol Biol **8**: 289.
Hey, J. (2005). "On the number of New World founders: a population genetic portrait of the peopling of the Americas." PLoS Biol **3**(6): e193.
Hey, J. (2010). "Isolation with migration models for more than two populations." Mol Biol Evol **27**(4): 905-920.
Ho, S. Y. and B. Shapiro (2011). "Skyline-plot methods for estimating demographic history from nucleotide sequences." Mol Ecol Resour **11**(3): 423-434.
Hoban, S., G. Bertorelle, et al. (2011). "Computer simulations: tools for population and evolutionary genetics." Nat Rev Genet **13**(2): 110-122.
Holliday, J. A., M. Yuen, et al. (2010). "Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests." Mol Ecol **19**(18): 3857-3864.
Ivy-Ochs, S., H. Kerschner, et al. (2008). "Chronology of the last glacial cycle in the European Alps." Journal Quaternary Science = JQS **23**(6-7): 559-573.
Janko, K., C. Marshall, et al. (2011). "Multilocus analyses of an Antarctic fish species flock (Teleostei, Notothenioidei, Trematominae): phylogenetic approach and test of the early-radiation event." Mol Phylogenet Evol **60**(3): 305-316.
Jombart, T., S. Devillard, et al. (2010). "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations." BMC Genet **11**: 94.
Joyce, P. and P. Marjoram (2008). "Approximately sufficient statistics and bayesian computation." Stat Appl Genet Mol Biol **7**(1): Article26.
Jukes, T. H. and C. R. Cantor (1969). Evolution of Protein Molecules, Academy Press.
Kenney, J. S., J. L. D. Smith, et al. (1995). "The Long-Term Effects of Tiger Poaching on Population Viability." Conservation Biology **9**(5): 1127-1133.
Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." Journal of Molecular Evolution **16**(2): 111-120.
Kingman, J. F. C. (1982). "The coalescent." Stochastic Processes and their Applications **13**(3): 235-248.
Kingman, J. F. C. (1982). "On the Genealogy of Large Populations." Journal of Applied Probability **19**: 27-43.
Kitchen, A., M. M. Miyamoto, et al. (2008). "Utility of DNA viruses for studying human host history: case study of JC virus." Mol Phylogenet Evol **46**(2): 673-682.
Kuo, L. and B. Mallick. (1998). "Variable Selection for Regression Models." 1. 60, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.9573.
La Mesa, M. and J. Ashford (2008). "Age and growth of ocellated icefish, <i>Chionodraco rastrospinosus</i> DeWitt and Hureau, 1979, from the South Shetland Islands." Polar Biology **31**(11): 1333-1342.
La Mesa, M. and M. Vacchi (2001). "Review - age and growth of high Antarctic notothenioid fish." Antarctic Science **13**: 227-235.
Leonard, J. A. (2008). "Ancient DNA applications for wildlife conservation." Mol Ecol **17**(19): 4186-4196.
Leonard, J. A., C. Vila, et al. (2005). "Legacy lost: genetic variability and population size of extirpated US grey wolves (Canis lupus)." Mol Ecol **14**(1): 9-17.
Leonard, J. A. and R. K. Wayne (2008). "Native Great Lakes wolves were not restored." Biol Lett **4**(1): 95-98.
Leuenberger, C. and D. Wegmann (2010). "Bayesian computation and model selection without likelihoods." Genetics **184**(1): 243-252.
Liao, P. C., D. C. Kuo, et al. (2010). "Historical spatial range expansion and a very recent bottleneck of Cinnamomum kanehirae Hay. (Lauraceae) in Taiwan inferred from nuclear genes." BMC Evol Biol **10**: 124.
Luikart, G., F. W. Allendorf, et al. (1998). "Distortion of allele frequency distributions provides a test for recent population bottlenecks." J Hered **89**(3): 238-247.
Magiorkinis, G., E. Magiorkinis, et al. (2009). "The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis." PLoS Med **6**(12): e1000198.
Manel, S., P. Berthier, et al. (2002). "Detecting Wildlife Poaching: Identifying the Origin of Individuals with Bayesian Assignment Tests and Multilocus Genotypes

Detección de la Caza Ilegal de Vida Silvestre: Identificación del Origen de Individuos con Pruebas de Asignación Bayesiana y Genotipos Multilocus." Conservation Biology **16**(3): 650-659.
Mank, J. E., J. E. Carlson, et al. (2004). "A Century of Hybridization: Decreasing Genetic Distance Between American Black Ducks and Mallards." Conservation Genetics **5**(3): 395-403.
Marjoram, P., J. Molitor, et al. (2003). "Markov chain Monte Carlo without likelihoods." Proc Natl Acad Sci U S A **100**(26): 15324-15328.
Marjoram, P. and J. D. Wall (2006). "Fast "coalescent" simulation." BMC Genet **7**: 16.
McVean, G. A. and N. J. Cardin (2005). "Approximating the coalescent with recombination." Philos Trans R Soc Lond B Biol Sci **360**(1459): 1387-1393.
Mech, L. D. and U. S. Seal (1987). "Premature Reproductive Activity in Wild Wolves." Journal of Mammalogy **68**(4): 871-873.
Miller, C. R. and L. P. Waits (2003). "The history of effective population size and genetic diversity in the Yellowstone grizzly (Ursus arctos): implications for conservation." Proc Natl Acad Sci U S A **100**(7): 4334-4339.
Mitchell, M. (1998). An Introduction to Genetic Algorithms. Cambridge, MA, USA, MIT Press.
Muñoz-Fuentes, V., A. Green, et al. (2005). "Population structure and loss of genetic diversity in the endangered white-headed duck, <i>Oxyura leucocephala</i&gt." Conservation Genetics **6**(6): 999-1015.
Munoz-Fuentes, V., C. Vila, et al. (2007). "Hybridization between white-headed ducks and introduced ruddy ducks in Spain." Mol Ecol **16**(3): 629-638.
Near, T. J., A. Dornburg, et al. (2011). "Climate change, antifreeze, and the evolutionary diversification of Antarctic fishes." Submitted.
Near, T. J., J. J. Pesavento, et al. (2004). "Phylogenetic investigations of Antarctic notothenioid fishes (Perciformes: Notothenioidei) using complete gene sequences of the mitochondrial encoded 16S rRNA." Mol Phylogenet Evol **32**(3): 881-891.
Nei, M. (1987). Molecular Evolutionary Genetics. New York, Columbia University Press.
Neuenschwander, S., C. R. Largiader, et al. (2008). "Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework." Mol Ecol **17**(3): 757-772.
Ng, K. K. S., S. L. Lee, et al. (2009). "Impact of selective logging on genetic diversity of two tropical tree species with contrasting breeding systems using direct comparison and simulation methods." Forest Ecology and Management **257**(1): 107-116.
Nichols, R. A., M. W. Bruford, et al. (2001). "Sustaining genetic variation in a small population: evidence from the Mauritius kestrel." Mol Ecol **10**(3): 593-602.
Nielsen, R. and J. Wakeley (2001). "Distinguishing migration from isolation: a Markov chain Monte Carlo approach." Genetics **158**(2): 885-896.
Nunes, M. A. and D. J. Balding (2010). "On optimal selection of summary statistics for approximate Bayesian computation." Stat Appl Genet Mol Biol **9**(1): Article34.
Padhukasahasram, B., P. Marjoram, et al. (2008). "Exploring population genetic models with recombination using efficient forward-time simulations." Genetics **178**(4): 2417-2427.
Papetti, C., E. Susana, et al. (2009). "Spatial and temporal boundaries to gene flow between Chaenocephalus aceratus populations at South Orkney and South Shetlands." Marine Ecology Progress Series **376**: 269-281.
Patarnello, T., S. Marcato, et al. (2003). "Phylogeography of the Chionodraco genus (Perciformes, Channichthydae) in the Southern Ocean." Mol Phylogenet Evol **28**(3): 420-429.
Peng, B. and M. Kimmel (2007). "Simulations provide support for the common disease-common variant hypothesis." Genetics **175**(2): 763-776.
Perry, W. L., D. M. Lodge, et al. (2002). "Importance of hybridization between indigenous and nonindigenous freshwater species: an overlooked threat to North American biodiversity." Syst Biol **51**(2): 255-275.
Pritchard, J. K., M. T. Seielstad, et al. (1999). "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." Mol Biol Evol **16**(12): 1791-1798.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.
Pybus, O. G., A. Rambaut, et al. (2000). "An integrated framework for the inference of viral population history from reconstructed genealogies." Genetics **155**(3): 1429-1437.
Ramakrishnan, U., E. A. Hadly, et al. (2005). "Detecting past population bottlenecks using temporal genetic data." Mol Ecol **14**(10): 2915-2922.
Rambaut, A., O. G. Pybus, et al. (2008). "The genomic and epidemiological dynamics of human influenza A virus." Nature **453**(7195): 615-619.
Ray, N. and L. Excoffier (2009). "Inferring past demography using spatially explicit population genetic models." Hum Biol **81**(2-3): 141-157.
Rodrigo, A. G. and J. Felsenstein (1999). Coalescent approaches to HIV population genetics. Molecular evolution of HIV. e. K. Crandall. Baltimore, Md., Johns Hopkins University Press**:** 233–272.
Rosenberg, N. A. and M. Nordborg (2002). "Genealogical trees, coalescent theory and the analysis of genetic polymorphisms." Nat Rev Genet **3**(5): 380-390.
Row, J. R., R. J. Brooks, et al. (2011). "Approximate Bayesian computation reveals the factors that influence genetic diversity and population structure of foxsnakes." J Evol Biol **24**(11): 2364-2377.
Saltonstall, K. (2002). "Cryptic invasion by a non-native genotype of the common reed, Phragmites australis, into North America." Proc Natl Acad Sci U S A **99**(4): 2445-2449.
Shapiro, B., A. J. Drummond, et al. (2004). "Rise and fall of the Beringian steppe bison." Science **306**(5701): 1561-1565.
Sjodin, P., A. E. Sjostrand, et al. (2012). "Resequencing data provide no evidence for a human bottleneck in Africa during the penultimate glacial period." Mol Biol Evol.
Soares, P., L. Ermini, et al. (2009). "Correcting for purifying selection: an improved human mitochondrial molecular clock." Am J Hum Genet **84**(6): 740-759.
Stelkens, R. B., K. A. Young, et al. (2010). "The accumulation of reproductive incompatibilities in African cichlid fish." Evolution **64**(3): 617-633.
Stiller, M., G. Baryshnikov, et al. (2010). "Withering away--25,000 years of genetic decline preceded cave bear extinction." Mol Biol Evol **27**(5): 975-978.
Strasburg, J. L. and L. H. Rieseberg (2010). "How robust are "isolation with migration" analyses to violations of the im model? A simulation study." Mol Biol Evol **27**(2): 297-310.
Strimmer, K. and O. G. Pybus (2001). "Exploring the demographic history of DNA sequences using the generalized skyline plot." Mol Biol Evol **18**(12): 2298-2305.
Tajima, F. (1989). "The effect of change in population size on DNA polymorphism." Genetics **123**(3): 597-601.
Tamura, K. and M. Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." Mol Biol Evol **10**(3): 512-526.
Tavaré, S. (1984). "Line-of-descent and genealogical processes, and their applications in population genetics models." Theor Popul Biol **26**(2): 119-164.
Tavaré, S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. American Mathematical Society: Lectures on Mathematics in the Life Sciences, Amer Mathematical Society. **17:** 57-86.
Tavare, S., D. J. Balding, et al. (1997). "Inferring coalescence times from DNA sequence data." Genetics **145**(2): 505-518.
Team, R. D. C. (2010). " R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria.
Templeton, A. R. (2009). "Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation." Mol Ecol **18**(2): 319-331.
Templeton, A. R. (2010). "Correcting approximate Bayesian computation." Trends Ecol Evol **25**(9): 488-489; author reply 490-481.
Thatje, S., C. D. Hillenbrand, et al. (2005). "On the origin of Antarctic marine benthic community structure." Trends Ecol Evol **20**(10): 534-540.

Valdiosera, C. E., J. L. Garcia-Garitagoitia, et al. (2008). "Surprising migration and population size dynamics in ancient Iberian brown bears (Ursus arctos)." Proc Natl Acad Sci U S A **105**(13): 5123-5128.
Veeramah, K. R., D. Wegmann, et al. (2011). "An Early Divergence of KhoeSan Ancestors from Those of Other Modern Humans Is Supported by an ABC-Based Analysis of Autosomal Resequencing Data." Mol Biol Evol.
Wandeler, P., P. E. Hoeck, et al. (2007). "Back to the future: museum specimens in population genetics." Trends Ecol Evol **22**(12): 634-642.
Wegmann, D., C. Leuenberger, et al. (2010). "ABCtoolbox: a versatile toolkit for approximate Bayesian computations." BMC Bioinformatics **11**: 116.
Weiss, G. and A. von Haeseler (1998). "Inference of population history using a likelihood approach." Genetics **149**(3): 1539-1546.
Wright, S. (1931). "Evolution in Mendelian Populations." Genetics **16**(2): 97-159.
Yang, Z. (1994). "Estimating the pattern of nucleotide substitution." Journal of Molecular Evolution **39**(1): 105-111.
Zharkikh, A. (1994). "Estimation of evolutionary distances between nucleotide sequences." J Mol Evol **39**(3): 315-329.