BMC Evolutionary Biology

Methodology article

Open Access

Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach

Michael J Hickerson*[1] and Christopher P Meyer[2]

Address: [1]Biology Department, Queens College, City University of New York, 65-30 Kissena Blvd, Flushing, NY 11367-1597, USA and [2]Smithsonian Institution, PO Box 37012, MRC 163, Washington, DC 20013-7012, USA

Email: Michael J Hickerson* - michael.hickerson@qc.cuny.edu; Christopher P Meyer - meyerc@si.edu

* Corresponding author

Published: 27 November 2008

BMC Evolutionary Biology 2008, **8**:322 doi:10.1186/1471-2148-8-322

Received: 28 May 2008
Accepted: 27 November 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/322

Abstract

Background: Marine allopatric speciation is an enigma because pelagic larval dispersal can potentially connect disjunct populations thereby preventing reproductive and morphological divergence. Here we present a new hierarchical approximate Bayesian computation model (HABC) that tests two hypotheses of marine allopatric speciation: 1.) "soft vicariance", where a speciation involves fragmentation of a large widespread ancestral species range that was previously connected by long distance gene flow; and 2.) peripatric colonization, where speciations in peripheral archipelagos emerge from sweepstakes colonizations from central source regions. The HABC approach analyzes all the phylogeographic datasets at once in order to make across taxon-pair inferences about biogeographic processes while explicitly allowing for uncertainty in the demographic differences within each taxon-pair. Our method uses comparative phylogeographic data that consists of single locus mtDNA sequences from multiple co-distributed taxa containing pairs of central and peripheral populations. We use the method on two comparative phylogeographic data sets consisting of cowrie gastropod endemics co-distributed in the Hawaiian (11 taxon-pairs) and Marquesan archipelagos (7 taxon-pairs).

Results: Given the Marquesan data, we find strong evidence of simultaneous colonization across all seven cowrie gastropod endemics co-distributed in the Marquesas. In contrast, the lower sample sizes in the Hawaiian data lead to greater uncertainty associated with the Hawaiian estimates. Although, the hyper-parameter estimates point to soft vicariance in a subset of the 11 Hawaiian taxon-pairs, the hyper-prior and hyper-posterior are too similar to make a definitive conclusion. Both results are not inconsistent with what is known about the geologic history of the archipelagos. Simulations verify that our method can successfully distinguish these two histories across a wide range of conditions given sufficient sampling.

Conclusion: Although soft vicariance and colonization are likely to produce similar genetic patterns when a single taxon-pair is used, our hierarchical Bayesian model can potentially detect if either history is a dominant process across co-distributed taxon-pairs. As comparative phylogeographic datasets grow to include > 100 co-distributed taxon-pairs, the HABC approach will be well suited to dissect temporal patterns in community assembly and evolution, thereby providing a bridge linking comparative phylogeography with community ecology.

Background

Allopatric speciation is an enigma in many marine organisms because larval dispersal can potentially connect disjoint populations and thereby prevent the reproductive and morphological divergence that arises from prolonged isolation [1-7]. This is especially enigmatic in the Indo-Pacific region where many species range freely across this expanse without evidence for barriers to genetic exchange. This marine region harbors the planet's highest species diversity and endemism of marine fauna, and with the absence of explicit barriers, some have pushed controversial models of sympatric speciation to explain this elevated diversity [8,9]. Even the proposed competing models of geographic speciation in the Indo-Pacific remain contentious and generally involve different facets of the classic dispersal or vicariance models for speciation. Under the one model (the "soft vicariance" model), speciations in peripheral archipelagoes result from a large widespread patchy ancestral species range connected by long distance gene flow that is eventually interrupted by oceanographic changes in temperature, sea level and/or currents [10-12] which leads to peripheral isolation and endemism. Under a second model (the "colonization" model), speciations in peripheral (i.e peripatric) archipelagoes emerge from sweepstakes centrifugal colonizations from high-diversity central areas followed by prolonged periods of isolation with potential inward range shifts towards the central region [8,9,13,14]. As in the case of terrestrial systems, using genetic data to distinguish these two scenarios is difficult because their expected genetic signatures are often similar, a situation that is exacerbated if demographic changes such as expansions and bottlenecks occur after an isolating event [10,15-17].

Likewise, discerning the modes of isolation and speciation using phylogenetic and phylogeographic data is often fraught with uncertainty because species can potentially shift their ranges [18] or loose the population genetic patterns associated with colonization >> 2N generations subsequent to isolation. Traditionally, vicariance and dispersal histories have been tested using phylogenetic approaches that use area cladograms [19,20], consensus methods [21], or parsimony [22] in combination with some method of ancestral character state reconstruction. Although many of these classic methods were biased to find vicariance, recent methods incorporate more complex biogeography histories [23,24] such as maximum likelihood [25,26] and Bayesian methods that use both distributional and phylogenetic data [27]. Likewise, empirical studies have increasingly found dispersal/colonization to be a more common force behind allopatric speciation [15,16,28-31]. Regardless, such analyses are often circular or ambiguous [32], and ancestral character reconstruction methods are always going to be hindered when elevated homoplasy in biogeographic patterns obscures the inferences in the older parts of a phylogeny [18,31,33].

Here we present an entirely different approach to testing for vicariance and dispersal histories. Instead of using phylogenetic comparative methods, we use coalescent population genetics to estimate ancestral demographic patterns across co-distributed taxa within a community. While this is in the spirit of previous suggested approaches that blend systematics and population genetics [34], the hierarchical Bayesian approach presented here tests vicariance and dispersal across taxon-pairs instead of doing so one at a time. Specifically we extend a hierarchical approximate Bayesian model (HABC) [35,36] in order to quantify the strength of these two alternative models of allopatric isolation across marine endemic taxa that are co-distributed in peripheral archipelagoes.

Using HABC allows sidestepping the requirement of an explicit likelihood function. Instead, it uses a probabilistic simulation model to generate data sets to compare with the empirical data. By using summary statistics, one can easily compare the simulated and empirical data in order to estimate parameters of the simulation model via an approximate sample of the posterior distribution. In HABC we use hyper-parameters that describe processes across co-distributed taxon-pairs as well as sub-parameters that describe the demographic history of each taxon-pair.

First we describe the population genetic models whose hyper-parameters we want to estimate, and then we describe HABC. After detailing the HABC model, the summary statistics and the HABC implementation, we test these two biogeographic hypotheses given two comparative phylogeographic datasets: mtDNA CO1 data collected from multiple cowrie gastropod species that are endemic to the Hawaiian and Marquesan archipelagos. Specifically we use HABC to test whether marine vicariance ("soft vicariance") (H_1) or colonization (H_2) is the dominant isolating mechanism in either of these two marine communities. After using HABC to choose the best model of community isolation, we then use HABC to estimate temporal congruence in soft vicariance and/or colonization. We specifically use mtDNA sequence data collected from each co-distributed peripheral endemic taxon as well as each of the respective sister species which are usually more geographically widespread (Additional file 1).

Although this method specifically addresses questions relevant to the species diversity and patterns of endemism in the Indo-Pacific, it will be broadly applicable to many comparative phylogeographic datasets and biogeographic settings.

Methods description

Soft vicariance and colonization

Rather than classical terrestrial vicariance where a large ancestral population is broken up into two isolated sister populations (Figure 1A), our "soft vicariance" scenario (H_1; Figure 1B) has two ancestral populations with effective sizes $(\theta_\tau)_1$ and $(\theta_\tau)_2$ that are connected by high to moderate gene flow (M_1 = 1.0 to 100.0 migrants per generation) until τ_V, when M_1 decreases to 0.0 – 1.0 migrants per generation (M_2). If this second period is prolonged, then effective isolation and divergence can occur [37]. At τ_V, the sizes of the two sister populations ($(\theta_\tau)_1$ and $(\theta_\tau)_2$) remain the same size or begin to grow exponentially until they reach their present sizes (θ_1 and θ_2) at $\tau = 0$ depending on the draw from the prior (Figure 1B). As in [16,17], time of vicariance, population sizes and migration rates are all free to vary across taxon-pairs according to their prior distributions.

Under the colonization scenario (H_2; Figure 1C), one of the sister populations is founded by a very small number of individuals $(\theta_\tau)_2$ that come from a larger source population $(\theta_\tau)_1$ at the time of colonization, τ_C with subsequent isolation. The small colonizing population $(\theta_\tau)_2$, then grows exponentially until it reaches its present effective size of θ_2 at $\tau = 0$. The primary parametric expectation that distinguishes marine vicariance (H_1) from colonization (H_2) is the relatively small effective population size of the colonized population ($(\theta_\tau)_2$) at the putative time of colonization (τ_C). Secondarily, the possibility of gene flow subsequent to the vicariance event τ_V further distinguishes H_1 and H_2, although this assumption can potentially be relaxed. With regards to different patterns in the molecular genetic data under H_1 and H_2, under colonization (H_2) samples from peripheral populations will likely accumulate a surplus of rare alleles due to having a current effective population size that greatly expanded from a small size after the colonization time τ_C. In addition, there is likely to be generally more genetic diversity under soft vicariance (H_1) due to there being two ancestral populations rather than just one.

To statistically quantify the relative support of these two hypotheses (H_1 and H_2) across Y co-distributed peripheral endemics and their sister taxa given DNA sequence data, we extend and modify the hierarchical approximate Bayesian computation (HABC) framework of [35,36]. We also use this framework to estimate the temporal congruence of both vicariance and colonization across the Y phylogeographic data sets.

Although one could independently estimate $(\theta_\tau)_2$ in each of the Y data sets and use all of the independent posterior densities of $(\theta_\tau)_2$ to measure the support of H_1 and H_2 across the Y pairs, implementation of a hierarchical model accomplishes this from a single analysis and uses more information from the data via "borrowing strength" [38-40].

Hierarchical approximate Bayesian computation

Our implementation of HABC is based on the framework presented in [35,36,41], and we review the important features here. In HABC, sub-parameters (Φ; within taxon-pair parameters) are conditional on "hyper-parameters" (ϕ) that quantify the variability of Φ among the Y taxon-pairs. Instead of explicitly calculating the likelihood expression P(Data|ϕ, Φ) to get a posterior distribution, we sample from the posterior distribution P((ϕ, Φ)|Data) by simulating the data K times under a coalescent model using candidate parameters randomly drawn from the joint hyper-prior and sub-prior distribution P(ϕ, Φ). A summary statistic vector $\mathbf{D}_i$ for each simulated dataset is then compared to the observed summary statistic vector $\mathbf{D}^*$ in order to generate random observations from the joint posterior distribution $f(\phi_i, \Phi_i|\mathbf{D}_i)$ by way of a rejection/acceptance algorithm followed by a weighted local linear regression step [42].

The rejection/acceptance algorithm involves calculating a summary statistic vector from the observed data and each of the K simulated data sets. Each simulated data set is generated using parameters that are randomly drawn from the joint prior. Following [35,42], K Euclidian distances between the normalized observed summary statistic vector $\mathbf{D}^*$ and each of the K normalized summary statistic vectors are then calculated ($||\mathbf{D}_i - \mathbf{D}^*|| = d$). An arbitrary proportion (tolerance) of the K simulations with the lowest d values are then used to obtain an approximate sample from the joint posterior after weighting and transforming the accepted parameter values using local linear regression [35,42]. After the local linear regression step, accepted and transformed parameter values that fall outside their respective prior bounds are subsequently transformed to have the values of their respective prior boundaries. For example, if an accepted parameter value with a uniform prior of [0.0,1.0] is transformed by local linear regression to a negative number, it is subsequently transformed to 0.0.

Hierarchical Model of Community Colonization and Vicariance

Model hyper-parameters and sub-parameters are listed and described in Additional file 2. Three hyper-parameters are drawn from their respective hyper-prior distributions (Additional file 2A), and these include: 1.) Z, the number of descendent populations per Y taxon-pairs that arise by colonization at times $T_C = \{\tau_C^1, ..., \tau_C^Z\}$; 2.) the number of *different* vicariance times $\bar{\Psi}_V = \{t_V^1, ..., t_V^{\Psi_V}\}$

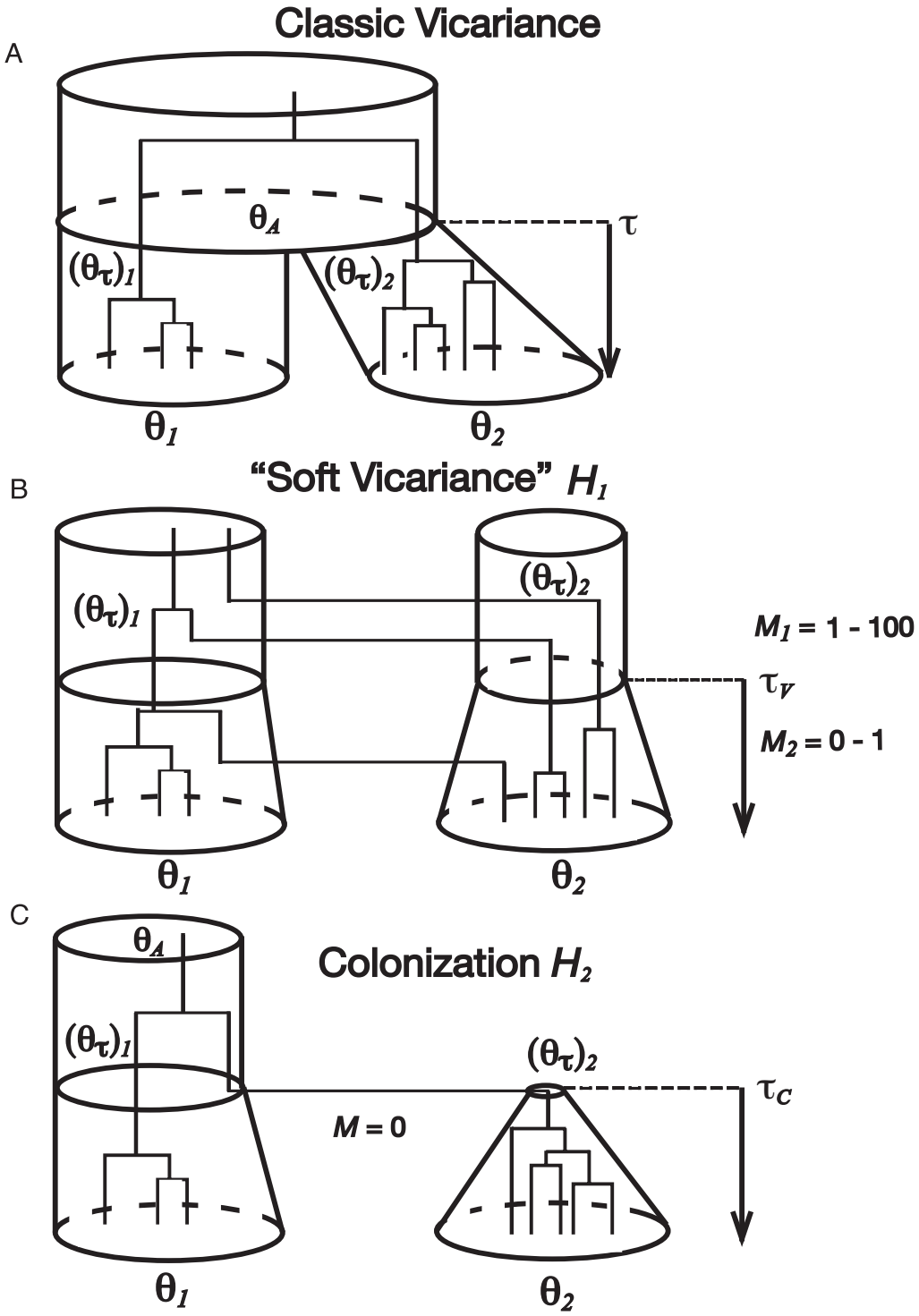

Figure 1

Three models of allopatric isolation. (A) Classic vicariance where a large ancestral population is broken up into two isolated sister populations. (B) Marine vicariance or "soft vicariance" where two ancestral populations with effective sizes $(\theta_\tau)_1$ and $(\theta_\tau)_2$ are connected by high to moderate gene flow (M_1 = 1.0 to 100.0 migrants per generation) until τ_V, when M_1 decreases to 0.0 – 1.0 migrants per generation (M_2). (C) Isolation by colonization, where one of the sister populations is founded by a very small number of individuals $(\theta_\tau)_2$ that come from a larger source population $(\theta_\tau)_1$ at the time of colonization, τ_C.

across (*Y-Z*) *actual* vicariance times $T_V = \{\tau_V^1, ..., \tau_V^{(Y-Z)}\}$ and 3.) Ψ_C, the number of *different* colonization times $\bar{\Psi}_C = \{t_C^1, ..., t_C^{\Psi_C}\}$ across *Z actual* colonization times $T_C = \{\tau_C^1, ..., \tau_C^Z\}$.

The *Z* colonized populations (H_2) and remaining (*Y-Z*) population-pairs that arise via vicariance (H_1) then draw their population mutation sub-parameters ($\vec{\theta} = \{\theta^1, ..., \theta^Y\}$, $\vec{\theta}_1 = \{\theta_1^1, ..., \theta_1^Y\}$, and $\vec{\theta}_2 = \{\theta_2^1, ..., \theta_2^Y\}$) from their respective sub-priors (Additional file 2C). Each taxon-pair's population mutation parameter θ^i is equal to the sum of θ_1^i and θ_2^i, the population mutation parameters of the descendent taxon-pairs at $\tau = 0$ (present time). In this case $\theta = 2N\mu$ ($2N$ is the sum of the two haploid effective female population sizes of each pair of descendent populations and μ is the per gene per generation mutation rate).

Subsequently, each of the *Y* taxon-pairs draw their remaining sub-parameters from two different sets of sub-priors (Additional file 2D) that differentially characterize the two different histories (H_1 and H_2). Importantly, the uniform sub-prior for $(\theta_\tau)_2$ is [0.0, 0.05] for each of the *Z* species-pairs that arose via colonization (H_2), but $(\theta_\tau)_2$ is drawn from the uniform sub-prior [0.0, 1.0] for each of the other (*Y - Z*) taxon-pairs that arose through vicariance (H_1). Additionally, two sets of migration sub-parameters ($\vec{M}_1 = \{M_1^1, ..., M_1^{(Y-Z)}\}$ and $\vec{M}_2 = \{M_2^1, ..., M_2^{(Y-Z)}\}$) are drawn from their respective sub-priors for the (*Y - Z*) taxon-pairs that arose through vicariance (H_1), whereas there is no migration under the colonization model (Figure 1C; Additional file 2D). In both vicariance and colonization models, the relative effective size of the ancestral central populations ($(\vec{\theta}_\tau)_1 = \{(\theta_\tau)_1^1, ..., (\theta_\tau)_1^{(Y-Z)}\}$ and $(\vec{\theta}_\tau)_1 = \{(\theta_\tau)_1^1, ..., (\theta_\tau)_1^Z\}$) are also drawn from their respective sub-priors [0.5, 1.0]. If $(Y - Z) \geq 1$, the Ψ_V *different* vicariance times ($\bar{\Psi}_V = \{t_V^1, ..., t_V^{\Psi_V}\}$) are drawn from the uniform prior [0.0, 5.0]. Likewise, if $Z \geq 1$, the Ψ_C *different* colonization times ($\bar{\Psi}_C = \{t_C^1, ..., t_C^{\Psi_C}\}$) are drawn from the uniform prior [0.0, 5.0]. After the $\bar{\Psi}_V$ *different* viciarance times are drawn, they are randomly assigned to the (*Y - Z*) taxon-pairs that arose through vicariance, such that the (*Y - Z*) *actual* vicariance times are $T_V = \{\tau_V^1, ..., \tau_V^{(Y-Z)}\}$. Specifically, the Ψ_V *different* vicariance times ($t_V^1, ..., t_V^{\Psi_V}$) are sequentially assigned to the first Ψ_V *actual* times $\tau_V^1, ..., \tau_V^{\Psi_V}$. The remaining actual times ($\tau_V^{(\Psi_V+1)}, ..., \tau_V^{(Y-Z)}$) are assigned by randomly drawing with replacement from the $\bar{\Psi}_V$ matrix of *different* times $\{t_V^1, ..., t_V^{\Psi_V}\}$. Likewise, the *actual* colonization times ($T_C = \{\tau_C^1, ..., \tau_C^Z\}$) are drawn using the same method (Additional file 2D). Both sets of *actual* vicariance and *actual* colonization times (T_V and T_C) are in units of θ/μ generations, where θ is each taxon-pair's population mutation parameter and μ is the per gene per generation mutation rate.

In addition to hyper-parameter estimation, we also use the HABC algorithm to sample from the posterior distributions of sub-parameter summaries (Additional file 2E) in order to quantify the support for H_1 and H_2 and secondarily estimate levels of temporal congruence in colonization and/or soft vicariance. Namely, we obtain estimates of the arithmetic means of three sub-parameters ($E((\theta_\tau)2)$, $E(\tau_C)$, $E(\tau_V)$), as well as the dispersion indexes of τ_C and τ_V (Ω_C and Ω_V respectively). The sub-parameter summary $E((\theta_\tau)_2)$ is expected to be < 0.05 if H_2 is dominate across the Y taxon-pairs (Z = Y). The dispersion indexes Ω_C and Ω_V of the *Z* colonization times ($\tau_C^1, ..., \tau_V^Z$) and the (*Y - Z*) vicariance times ($\tau_V^1, ..., \tau_V^{(Y-Z)}$) measure the ratio of the variance to the mean of these two sets of times and are therefore expected to be ≈ 0.0 when if there is temporal congruence in colonization or soft vicariance.

Two Stage Model Implementation

We implement the hierarchical analysis in two stages. In stage 1, an unconstrained general model is used to quantify the support for H_1 and H_2 across the *Y* taxon pairs. This first stage is accomplished by simulating *K* random draws from a general hyper-prior and using the HABC algorithm to sample from the posteriors of? *Z* and $E((\theta_\tau)_2)$. Under our hierarchical model, H_1 and H_2 are equally probable because *Z* is a hyper-parameter drawn from the discrete uniform hyper-prior distribution [0, *Y*].

In the stage 2 analysis, *K* random draws are taken from a constrained hyper-prior where *Z* is fixed to be the mode of its posterior distribution obtained in stage 1. There are equal numbers of hyper-prior draws (*K*) obtained in both

stages. The stage 2 analysis allows obtaining posterior samples of hyper-parameters and sub-parameter summaries (Additional file 2B &2E) that quantify and summarize the levels of temporal concordance in colonization ($T_C = \{\tau_C^1, ..., \tau_C^Z\}$) and/or vicariance ($T_V = \{\tau_V^1, ..., \tau_V^{(Y-Z)}\}$). These include: 1.) Ψ_C, the number of *different* colonization times per Z colonization events; 2.) Ψ_V, the number of *different* vicariance times per (Y - Z) vicariance events; 3.) Ω_C, the dispersion index of τ_C (the ratio of the variance to the mean in the Z *actual* colonization times, T_C); and 4.) Ω_V, the dispersion index of τ_V (the ratio of the variance to the mean in (Y - Z) *actual* vicariance times, T_V).

Summary Statistic Vector for HABC Acceptance/Rejection Algorithm

In order to implement the HABC procedure, we use two modified versions of the summary statistic vector **D** used in [35]. For the Marquesas summary statistic vector ($\mathbf{D}_{Marquesas}$), we calculate eight summary statistics collected from each taxon-pair (56 total). This includes π_b (average pairwise differences between each central and peripheral Marquasan taxon-pair), π (average pairwise differences among all individuals within each taxon-pair), π_w (average pairwise differences within descendent populations of each taxon-pair), θ_W (Watterson's estimator of θ of each taxon-pair), and Var(π - θ_W). For π, θ_W and Var(π - θ_W), each of the Y taxon-pairs are treated as a single population sample ($\vec{n}_1 + \vec{n}_2$). The Marquesas summary statistic vector, $\mathbf{D}_{Marquesas}$ also includes calculating π, θ_W, and Var(π - θ_W) in each of the Y peripheral Marquasan samples ($\vec{n}_2$) that are putatively colonized and we denote these as π_2, $(\theta_W)_2$, and Var(π - θ_W)$_2$. Under this scheme, the vector $\mathbf{D}_{Marquesas}$ is

$$\mathbf{D}_{Marquesas} = \begin{Bmatrix} \pi^1 & \pi_w^1 & \theta_W^1 & \mathrm{Var}(\pi - \theta_W)^1 & \pi_2^1 & (\theta_W)_2^1 & \mathrm{Var}(\pi - \theta_W)_2^1 & \pi_b^1 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ \pi^Y & \pi_w^Y & \theta_W^Y & \mathrm{Var}(\pi - \theta_W)^Y & \pi_2^Y & (\theta_W)_2^Y & \mathrm{Var}(\pi - \theta_W)_2^Y & \pi_b^Y \end{Bmatrix},$$

where each of the Y rows correspond to the Y taxon-pairs (Y = 7) and the eight columns correspond to the eight summary statistic classes. After these 8 × Y summary statistics are calculated, we must choose a way to consistently order the Y rows within $\mathbf{D}_{Marquesas}$. Instead of consistently ordering the rows by each taxon-pair's sample size, we increase the efficiency of our HABC estimator by ordering the rows based on the taxon-pair's Tajima's D [43] calculated from the taxon-pair's peripheral population sample ($\vec{n}_2$). For example, row 1 would contain the eight summary statistics collected from the taxon-pair with the lowest Tajima's D, and the Y^{th} row would contain the eight summary statistics collected from the taxon-pair with the highest Tajima's D.

The motive for this ordering procedure is to extract more information from the data with respect to the estimated hyper-parameters than would be accomplished by ordering consistently by sample size [35]. For an efficient HABC estimator, there should be a strong correlation between pair-wise differences in hyper-parameter values (i.e. E$(\theta_\tau)_2$ or Z) and Euclidian distances between corresponding pairs of summary statistic vectors from corresponding pairs of simulated data sets. If ordering by sample size rather than ranked values of an informative summary statistic, pair-wise values of Z *or* E$(\theta_\tau)_2$ are not predicted to correlate with Euclidian distances of **D** calculated from corresponding pairwise simulated data sets. This is because sample size has no bearing on how each of the Y taxon-pairs are assigned to histories H_1 and H_2 when drawing values of Z from the hyper-prior. On the other hand, ordering by an informative summary statistic will minimize Euclidian distances among data sets with equal or similar values of Z regardless of which of the Y taxon pairs were assigned histories H_1 and H_2. The consequent improved accuracy in HABC estimation that results from this ordering procedure is based on the *exchangeability* of the Y rows within $\mathbf{D}_{Marquesa}$ ($\mathbf{D}_1$,...,$\mathbf{D}_Y$). If Φ_i and $\mathbf{D}_i$ are invariant to the permutations of the indexes (1 ,..., Y) and the i^{th} taxon-pair's sample size is unrelated to the expectation of its Φ_i or $\mathbf{D}_i$, there is *exchangeability* in the model [44]. We order by the peripheral Tajima's D because it is a summary statistic that is predicted to be informative with respect to demographic parameter differences between histories H_1 and H_2 (i.e. the ratio of each taxon-pair's $(\theta_\tau)_2$ and θ_2).

For the analysis of the 11 Hawaiian taxon-pairs, we use a reduced summary statistic vector $\mathbf{D}_{Hawaii}$ to avoid null values that would arise in population samples that only included one individual (Additional file 1B). The vector $\mathbf{D}_{Hawaii}$ in this case is

$$\mathbf{D}_{Hawaii} = \begin{Bmatrix} \pi^1 & \theta_W^1 & \mathrm{Var}(\pi - \theta_W)^1 & \pi_b^1 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ (\pi)^Y & \theta_W^Y & \mathrm{Var}(\pi - \theta_W)^Y & \pi_b^Y \end{Bmatrix},$$

and the rows were likewise sorted by Tajima's D calculated from the peripheral populations.

Application on real data sets

Two Comparative Phylogeographic Implementations

We use this HABC method on two comparative phylogeographic data sets of Pacific cowrie gastropods (Cypraeidae). The first dataset consists of seven sister taxon-pairs of cowries that each consist of a descendent endemic species or sub-species that is distributed within the peripherally located Marquesas archipelago as well each sister taxon that is more geographically widespread (Additional file 1A). The second dataset consists of eleven sister taxon-pairs of cowries co-distributed within the peripherally located Hawaiian archipelago. Like in the former data set, each pair consists of a Hawaiian endemic and a more widespread sister (Additional file 1B). Both data sets consisted of 614 base pairs of the CO1 mtDNA locus collected from 2–93 individuals per taxon-pair (Additional file 1). The HABC procedure was implemented using a modified version of the MSBAYES comparative phylogeographic software pipeline [35,36] consisting of several C and R programs that are run with a Perl "front-end" and utilizes a finite sites version of Hudson's classic coalescent simulator [45]. For both analyses, 2,000,000 random draws were sampled from the hyper-prior and 1,000 – 2,000 accepted draws were used to construct hyper-posterior samples (tolerance of 0.0005 and 0.001 respectively) using the HABC acceptance/rejection algorithm.

The prior bounds for hyper-parameters and sub-parameters are given in Additional file 2. To explore the sensitivity of using different prior assumptions, we used two different upper bounds of θ (θ_{MAX} = 25.0 and 50.0 for the Marquesas data; θ_{MAX} = 50.0 and 100.0 for the Hawaiian data). These values correspond to 2× to 4× the range of within species θ estimates, where the average number of pairwise differences was used as an estimator of each species specific θ [46]. To further explore how sensitive results are to model assumptions, we alternatively ran the stage 1 analysis with the post-colonization migration prior allowed to be [0.0, 1.0] instead of zero under the colonization model (H_2), as well as allowing post-isolation migration (M_2) to be zero under the soft vicariance model (H_1).

We calculate Bayes factors to compare the relative hyper-posterior support of either history (soft vicariance and colonization) being dominate across all Y taxon pairs (e.g $Z = 0$ or $Z = Y$) against all other scenarios including mixed scenarios. We accomplish this by comparing relative hyper-posterior support of these two scenarios while accounting for the relative hyper-prior support for these two scenarios [47]. To calculate this Bayes factor, we use an arbitrary partition of hyper-parameter space to delineate where H_1 or H_2 is dominate across all Y taxon-pairs. For example, to evaluate the evidence of colonization being dominate across all Y taxon pairs ($Z = Y$) against all other scenarios ($Z < Y$), the approximate Bayes factor $B(Z = Y, Z < Y)$ is the ratio of the two approximate hyper-posteriors of these two scenarios divided by the ratio of the two hyper-priors of these two scenarios,

$$B(Z = Y, Z < Y) = (P(Z = Y|\mathbf{D} = \mathbf{D}^*)/P(Z < Y|\mathbf{D} = \mathbf{D}^*))/(P(Z = Y)/P(Z < Y))$$

Alternately, we examine these two scenarios by using an arbitrary partition of $E((\theta_\tau)_2)$ such that $E((\theta_\tau)_2) = 0.05$ represents a scenario where colonization is dominant across all Y taxon pairs, and $E((\theta_\tau)_2) > 0.05$ represents all other scenarios. In this case, the approximate Bayes factor is

$$B(E((\theta_\tau)_2) \le 0.05, E((\theta_\tau)_2) > 0.05) = \frac{(P(E((\theta_\tau)_2) \le 0.05 | \mathbf{D}=\mathbf{D}^*) / P(E((\theta_\tau)_2) > 0.05 | \mathbf{D}=\mathbf{D}^*))}{(P(E((\theta_\tau)_2) \le 0.05) / P(E((\theta_\tau)_2) > 0.05))}.$$

Evaluating the evidence of soft vicariance being dominant across all Y taxon-pairs ($Z = 0$) against all other scenarios ($Z > 0$) is identically accomplished by calculating the two corresponding Bayes factors $B(Z = 0, Z > 0)$ and $B(E((\theta_\tau)_2) > 0.05, E((\theta_\tau)_2) \le 0.05)$. To calculate each Bayes factor, we use the accepted hyper-parameter values from the hyper-posterior sample and the random draws from the hyper-prior.

Marquesas Results

The HABC analysis yielded a hyper-posterior that strongly supports H_2, where colonization histories (Figures 2A &2B) are inferred across all seven Marquesan endemics. Irrespective of tolerance (0.0005 or 0.001), prior assumptions on the upper bound of θ, (θ_{MAX}), the hyper-parameter mode estimates of Z and $E((\theta_\tau)_2)$ were 7.00 and 0.00 – 0.07 respectively (Figures 2A &2B; Additional file 3A &3B). Furthermore, when using the raw untransformed accepted values to obtain our posterior sample of Z, a history of community colonization is also inferred (mode estimate of Z is 7.0) albeit the 95% credibility intervals were wider. This suggests that the dominant mode of speciation has generally been from colonization by a small number of individuals followed by at least a 20-fold demographic expansion to current population levels. Both Bayes factors evaluating the evidence for colonization across all $Y = 7$ taxon-pairs against all other histories ($Z < 7$) indicate moderate support for the former scenario ($B(Z = Y, Z < Y) = 6.23$; $B(E((\theta_\tau)_2) > 0.05, E((\theta_\tau)_2) \le 0.05) = 6.79$).

In stage 2, where we constrained the hyper-parameter Z to be 7, the hyper-posterior best supported temporal concordance in colonization across all seven Marquesas

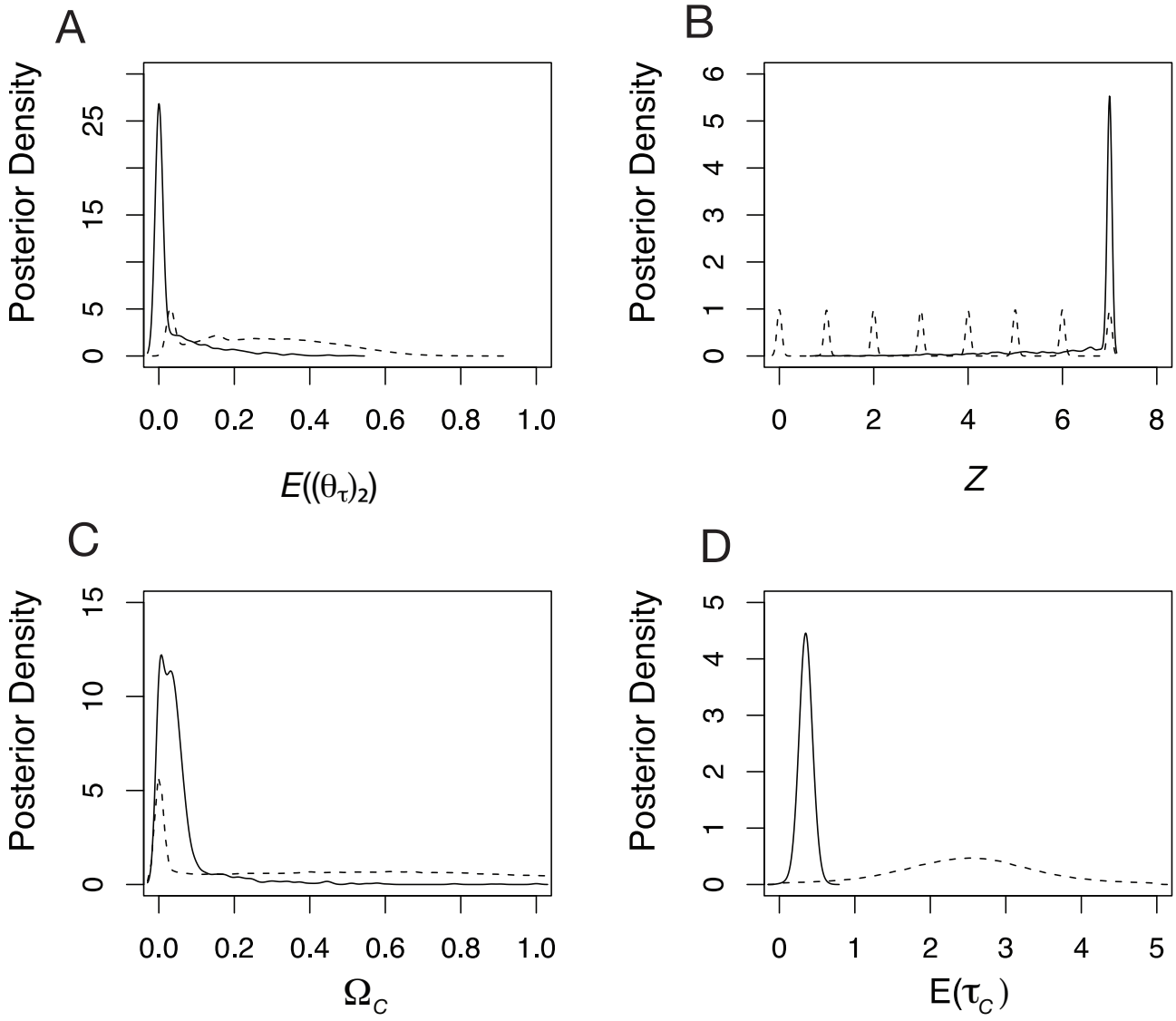

Figure 2

Hyper-posterior and Hyper-prior samples given Marquesas DNA sequence data (seven taxon-pairs). Dashed lines depict hyper-prior distributions and solid lines depict hyper-posterior distributions. **Stage 1 hyper-parameter estimates**: (A) Average effective population size $E((\theta_\tau)_2)$ of the putatively colonized endemic populations at the seven isolation times (T_C and T_V). (B) The number of descendent populations per seven (*Y*) taxon-pairs that arise by colonization at times $T_C = \{\tau_C^1, ..., \tau_C^Z\}$. **Stage 2 hyper-parameter estimates where hyper-prior is conditional given constant value of *Z* = 7**: (C) Ω_C,dispersion index of *Z* = 7 colonization times; Ω_C = Var(τ_C)/E(τ_C) where τ_C is colonization time. (D) E(τ_C), average colonization time across *Z* = 7 colonization times. For each estimate, tolerance was 0.001 (2,000 accepted draws) using the local regression algorithm.

endemics. In this case, estimates of Ψ_C and Ω_C were 1.00 (95% quantiles: 1.00 – 2.14) and 0.00 (95% quantiles: 0.00 – 0.17) respectively (Additional file 3A; Figure 2). The mean time of colonization E(τ_C) was 1.58 My ago if we assume a 1% divergence rate per My (Figure 2D). The Bayes factor evaluating the evidence for simultaneous colonization ($\Omega_C < 0.05$) against non-simultaneous colonization ($\Omega_C = 0.05$) yielded strong support for simultaneous colonization ($B(\Omega_C < 0.05, \Omega_C = 0.05) = 36.73$). In this case the Bayes factor is calculated from an

arbitrary partition of hyper-posterior space conditional on $Z = 7$. In this case the arbitrary threshold of simultaneous colonization is $\Omega_C < 0.05$ such that the Bayes factor is

$$B(\Omega_C < 0.05, \Omega_C \geq 0.05) = (P(\Omega_C < 0.05|\mathbf{D} = \mathbf{D}^*)/ P(\Omega_C \geq 0.05|\mathbf{D} = \mathbf{D}^*))/(P(\Omega_C < 0.05)/P(\Omega_C \geq 0.05)).$$

Furthermore, the hyper-posterior estimates were not sensitive to assumptions about post-isolation migration. Under the stage 1 analysis, estimates of Z and $E((\theta_\tau)_2)$ were 6.99 and 0.02 (95% credibility intervals of 3.34 – 7.00 and 0.00 – 0.32) when the post-colonization migration prior was [0.0, 1.0] instead of 0.0 under the colonization model (H_2).

Hawaii Results

In contrast to the Marquesas analysis, the hyper-posteriors were much more similar to the hyper-priors (Figure 3). The less informative posteriors are consistent with the lower Hawaiian sample sizes (Additional file 1B) and consequence of using of a reduced number of summary statistics ($\mathbf{D}_{Hawaii}$) for the HABC acceptance/rejection algorithm. Although the hyper-posterior of Z given the Hawaiian data suggests a mixed history of both colonization and soft vicariance (Figure 3), larger sample sizes will be required to verify this. This weak inference is demonstrated by Bayes factors giving weak support for vicariance or colonization across all 11 Hawaiian endemics. Specifically, the calculated Bayes factors $B(Z = 0, Z > 0)$, $B(Z = 11, Z < 11)$, $B(E((\theta_\tau)_2) < 0.05, (E((\theta_\tau)_2) = 0.05) < 1.0)$ and $B(E((\theta_\tau)_2) = 0.05, E((\theta_\tau)_2) < 0.05) < 1.0)$ where all weak (< 1.0). Although the Hawaiian data yielded weak inference, the credibility intervals for Z ranged from 0.00 to 9.22 (Additional file 3C and 3D), suggesting that the history of isolation likely involved both soft vicariance and colonization. Likewise, estimates of Ω_C and Ω_V did not suggest simultaneous vicariance or colonization, and likewise yielded less informative posteriors than obtained in the Marquesas analysis (Figures 3C &3D). As was the case of the Marquesas analysis, hyper-posterior estimates were not sensitive to tolerance, prior assumptions on the upper bound of θ, (Additional file 3C &3D), and prior assumptions of post-isolation migration. Under the stage 1 analysis using the alternative model assumptions where post-vicariance migration (M_2) was 0.0 under the soft vicariance model, estimates of Z and $E((\theta_\tau)_2)$ were respectively 3.64 and 0.66 (95% credibility intervals 0.00 – 9.74 and 0.35 – 0.97).

Simulation testing

Simulated Data Sets

One of the chief advantages of HABC and ABC methods is the ease at which one can evaluate the performance, bias, and precision of the estimator via simulations. Specifically, we can simulate pseudo-observed data sets with known hyper-parameter values and compare estimates with their true values. Even though the most time-consuming task is to simulate a large enough prior and/or hyper-prior in ABC and HABC, once it is produced for the analysis of the real observed data set, it can subsequently be used to quantify bias and precision on the pseudo-observed data sets.

To this end, we simulate pseudo-observed data sets with known values of Z and $E((\theta_\tau)_2)$ under the general model (stage 1), and Ψ_C, Ψ_V, $E(\tau_C)$, $E(\tau_V)$, Ω_C, and Ω_V under the constrained model (stage 2). We repeat this for both sample sizes used in the empirical implementation ($\mathbf{D}_{Marquesas}$ and $\mathbf{D}_{Hawaii}$). To simulate a data set (pseudo-observed data), all hyper-parameters and sub-parameters were randomly drawn from the prior and a corresponding $\mathbf{D}_{Marquesas}$ or $\mathbf{D}_{Hawaii}$ was subsequently calculated. For each corresponding pseudo-observed $\mathbf{D}_{Marquesas}$ and $\mathbf{D}_{Hawaii}$, a hyper-posterior sample was obtained from the HABC rejection-sampling algorithm given 2,000,000 random draws from the hyper-prior. For every pseudo-observed (simulated) data set, an estimate is obtained from the posterior mode. We report estimates using a tolerance of 0.001 corresponding to 2,000 accepted draws from the hyper-prior. For both $\mathbf{D}_{Marquesas}$ and $\mathbf{D}_{Hawaii}$ and stage 1 and two, 250 estimates were made from 250 simulated pseudo-observed datasets.

In addition to these evaluations, we assess the ability to distinguish H_1 and H_2 when the true history is a special asymmetrical case of soft vicariance (H_1). To this end, we obtain HABC estimates of Z and $E((\theta_\tau)_2)$ on 250 pseudo-observed datasets of size $\mathbf{D}_{Marquesas}$ simulated under H_1 where true values of Z and $E((\theta_\tau)_2)$ are 0 and 0.0 – 0.05 respectively. Estimates for Z and $E((\theta_\tau)_2)$ were obtained using a general prior (stage 1).

Another factor we explored was three other methods of post-acceptance transformation other than local linear regression (LLR) for obtaining a sample of the hyper-posterior distribution of the Z hyper-parameter. Because Z is a discrete integer (ranging from 0 to Y), it might be most appropriate to preserve Z as a discrete integer when using an ABC regression technique that implements polychotomous logit regression (PLR) [48-50]. We therefore used simulations to compare the effectiveness of LLR with: 1. PLR; 2. using the raw accepted values (RAW); and 3. a cumulative logit regression model (CLR). To compare the estimator bias and precision of these four methods, we calculated the root mean square error (RMSE) from 1000 estimates using each of these four methods on 1000 simulated pseudo-observed data sets with parameters randomly drawn from the priors. The methods for CLR and

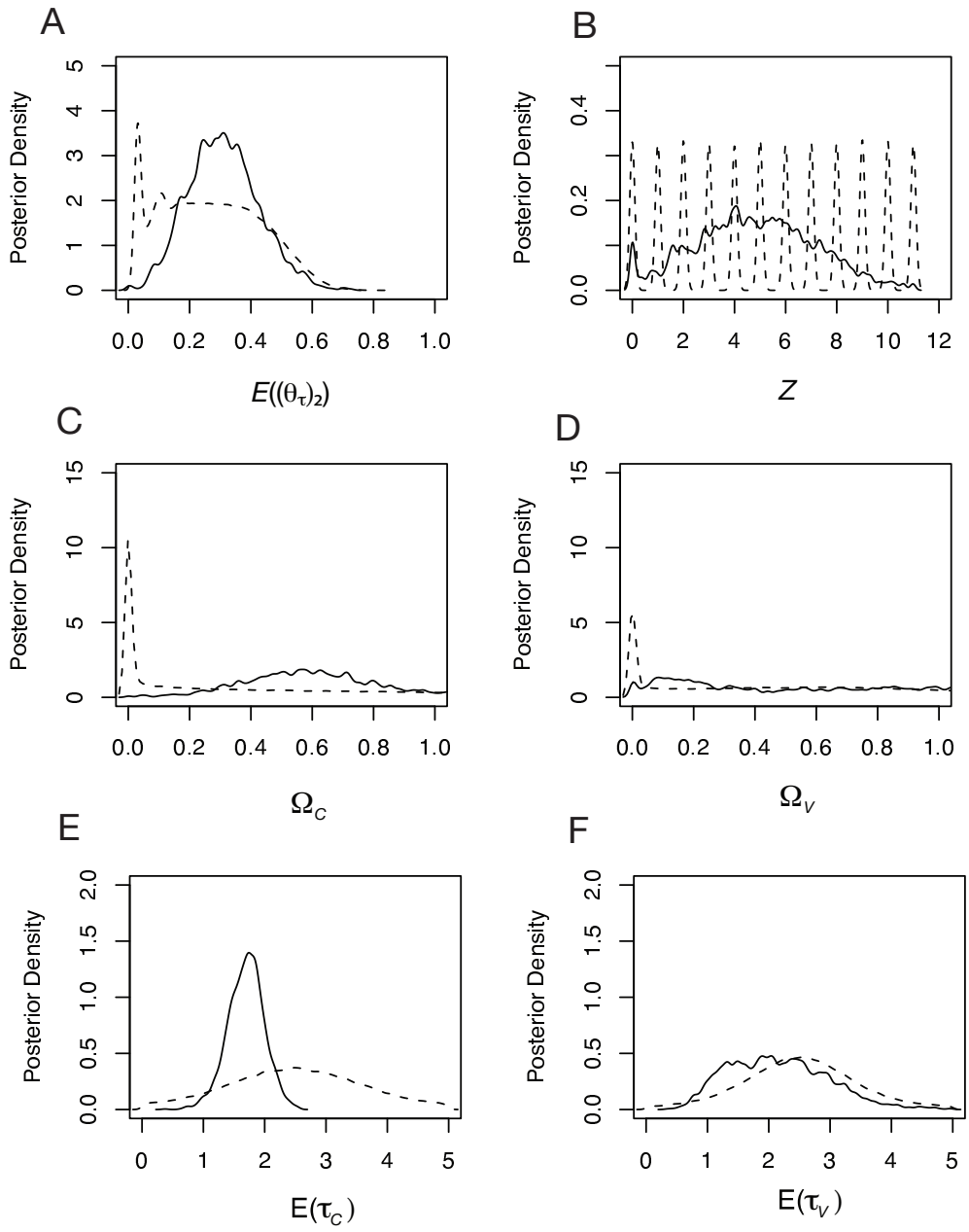

Figure 3

Hyper-posterior and Hyper-prior samples given Hawaiian DNA sequence data (11 taxon-pairs). Dashed lines depict hyper-prior distributions and solid lines depict hyper-posterior distributions. **Stage 1 hyper-parameter estimates**: (A) Average effective population size $E((\theta_\tau)_2)$ of the putatively colonized endemic populations at times the 11 isolation times (T_C and T_V) across the seven (Y) taxon-pairs. (B) The number of descendent populations per seven (Y) taxon-pairs that arise by colonization at times $T_C = \{\tau_C^1, ..., \tau_C^Z\}$. **Stage 2 hyper-parameter estimates where hyper-prior is conditional given constant value of *Z* = 4**: (C) Ω_C,dispersion index of $Z = 4$ colonization times; $\Omega_C = \text{Var}(\tau_C)/\text{E}(\tau_C)$ where τ_C is colonization time. (D) Ω_V,dispersion index of $Z = 7$ soft vicariance times; $\Omega_V = \text{Var}(\tau_V)/\text{E}(\tau_V)$ where τ_V is soft vicariance time time. (E) $\text{E}(\tau_C)$, average colonization time across $Z = 4$ colonization times. (F) $\text{E}(\tau_V)$, average soft vicariance time across 7 soft vicariance times. For each estimate, tolerance was 0.001 (2,000 accepted draws) using the local regression algorithm.

PLR are implemented in the VGAM package distributed by T. Yee under R http://www.stat.aukland.ac.nz/~yee. The LLR method is implemented from R functions made available by M. Beaumont.

Results of Simulation Testing

Our simulations identified conditions under which the HABC estimator is reliable as well as conditions under which it is less reliable. At stage 1 of the HABC analysis, the estimates of $E((\theta_\tau)_2)$ were consistently close to the corresponding true values of $E((\theta_\tau)_2)$, although the larger sample sizes matching the Marquesas data set ($\mathbf{D}_{Marquesas}$) yielded more accurate estimates of $E((\theta_\tau)_2)$ than using sample sizes matching the smaller Hawaii data set ($\mathbf{D}_{Hawaii}$; Figures 4A &4B). On the other hand, estimates of Z were less accurate, yet using $\mathbf{D}_{Marquesas}$ resulted in more accurate estimates of Z than when using $\mathbf{D}_{Hawaii}$ (Figures 4C &4D). Additionally, estimates of Z representing colonization across all Y taxon-pairs never resulted in false positives given $\mathbf{D}_{Marquesas}$ (Figure 4C). Specifically, when

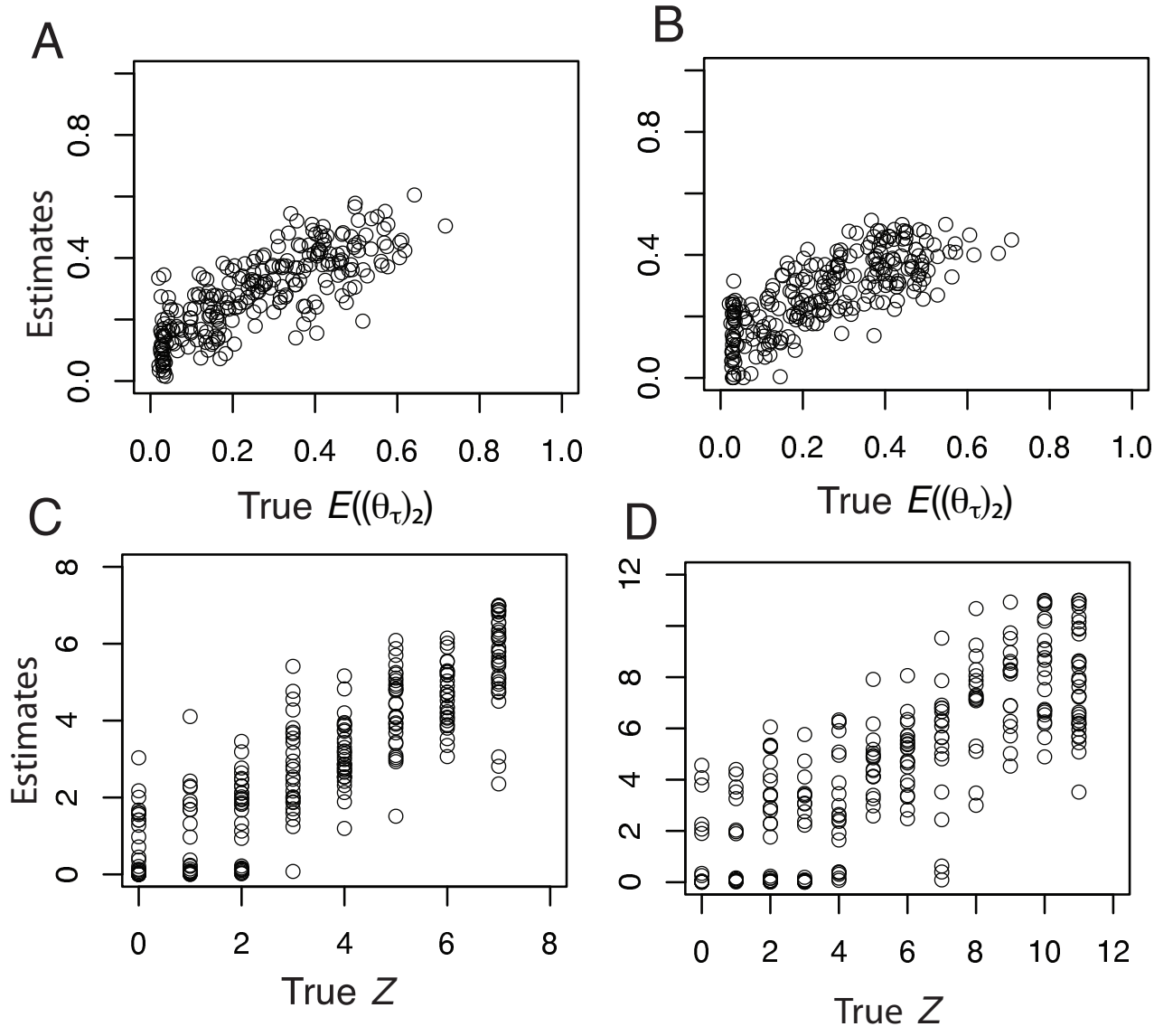

Figure 4

Estimator Performance: 250 true hyper-parameter values plotted against their posterior mode estimates (stage 1 model). Panels (A) and (C) are given samples sizes that are identical to the Marquesas sample sizes. Panels (B) and (D) are given samples sizes that are identical to the Hawaiian sample sizes. For each estimate, tolerance was 0.001 (2,000 accepted draws) using the local linear regression algorithm.

estimates of Z were 6 or 7, true Z values were always 6 or 7. Conversely, lower true values of Z yielded less accurate estimates of Z (Figures 4C &4D). For example, when estimates of Z were 0, true values ranged from 0 – 4 given $\mathbf{D}_{Marquesas}$ (Figure 4C). In general, the simulation analysis verified the statistical confidence in the empirical inference of colonization across all seven Marquesas endemics, yet demonstrated there to be more uncertainty in the inferred history of the 11 Hawaiian endemics.

Under the constrained stage 2 model, the simulations revealed a strategy for estimating the variability in the Z colonization times. Specifically, the best strategy would be to use estimates of Ω_C rather than Ψ_C (Figure 5). However, there was less accuracy and precision in the Ω_C estimates given $\mathbf{D}_{Hawaii}$ (Figure 5E). Unlike estimates of Ω_C, estimates of Ω_V or $E(\tau_V)$ were not as reliable (Figure 5H), perhaps due to the very small amount of migration (0 to 1.0 migrants per generation) after each "soft vicariance" event (τ_V) under H_1.

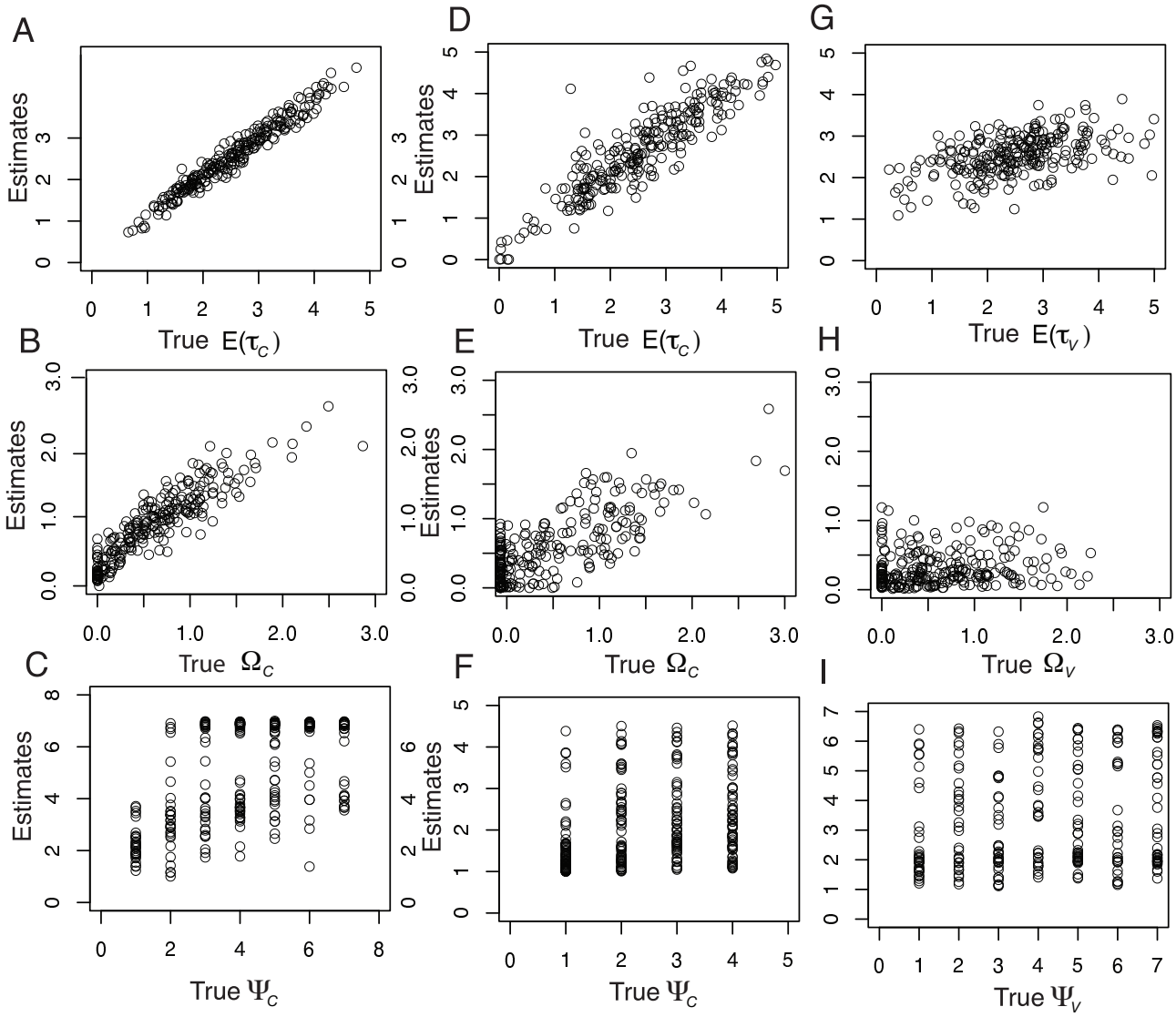

Figure 5

Estimator Performance: 250 true hyper-parameter values plotted against their posterior mode estimates (stage 2 model). Panels (A), (B) and (C) are given samples sizes that are identical to the Marquesas sample sizes and are obtained with the hyper-prior of Z fixed at 7 (stage 2 model). Panels (D) through (I) are given samples sizes that are identical to the Hawaiian sample sizes and are obtained with the hyper-prior of Z fixed at 4 (stage 2 model). For each estimate, tolerance was 0.001 (2,000 accepted draws) using the local linear regression algorithm.

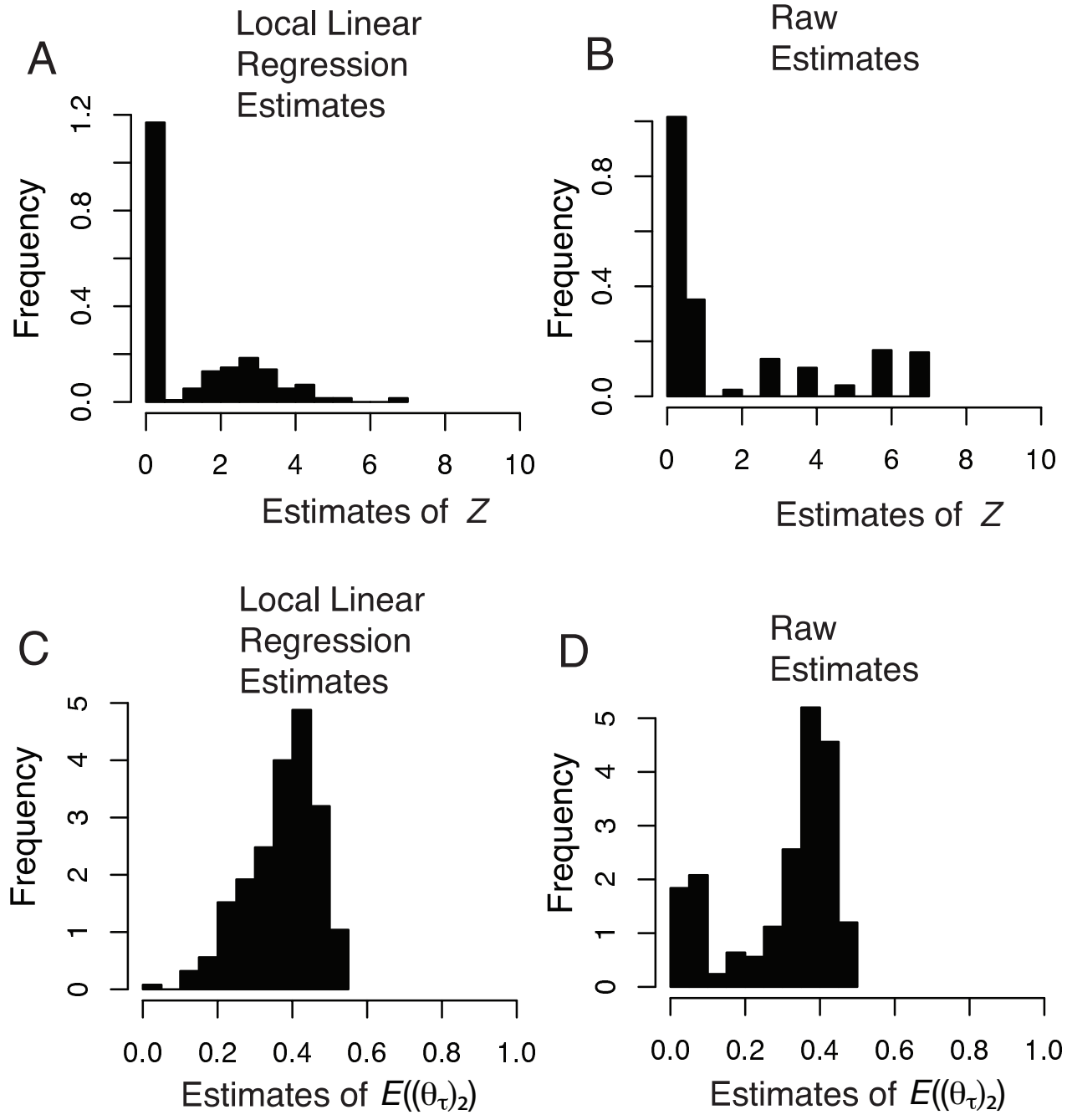

Figure 6

Estimator performance under constrained history of asymmetrical soft vicariance (True *Z* = 0; True $E((\theta_\tau)_2) < 0.05$; stage 1 model). Each panel depicts a frequency histogram of 250 mode estimates of *Z* (A and B) and $E((\theta_\tau)_2)$ (C and D) under a constrained history of asymmetrical soft vicariance where *Z* = 0 and $E((\theta_\tau)_2) < 0.05$. For all mode estimates, tolerance was 0.001 (2,000 accepted draws). Mode estimates for panels (A) and (C) are obtained using the local linear regression algorithm, whereas the mode estimates for panels (B) and (D) are obtained using the raw 2,000 accepted values.

The simulations also revealed that we have the ability to distinguish H_1 and H_2 when the true history is a special asymmetrical case of vicariance (H_1) where $Z = 0$ and each $(\theta_\tau)_2$ ranges from 0.0 to 0.05 under the $\mathbf{D}_{Marquesas}$ sample size configuration (Figure 6). In this case, 63% of the Z estimates were ≤ 1 (Figure 6A &6B). Likewise, estimates of $E((\theta_\tau)_2)$ were a reliable indicator of detecting H_1 under this special asymmetrical case of vicariance (H_1). Even though true values of $E((\theta_\tau)_2)$ ranged from 0.0 to 0.05, estimates of $E((\theta_\tau)_2)$ in this case were upwardly bias, with 90% of the $E((\theta_\tau)_2)$ estimates ranging from 0.21 to 0.48 (Figure 6C). Even though the $E((\theta_\tau)_2)$ estimator is upwardly biased given this special case, it is upwardly biased in the direction of correct inference of H_1 if one uses the criterion of $E((\theta_\tau)_2) >> 0.05$ to distinguish H_1 from H_2. In this case, our empirical estimates of Z and $E((\theta_\tau)_2)$ given the Marquesas data were not likely the result of extreme asymmetrical soft vicariance.

The root mean square error (RMSE) from the simulation study also revealed local linear regression (LLR) to outperform the three other methods that all keep the accepted values of Z as discrete integers (LLR RMSE = 1.32; RAW RMSE = 1.90; CLR RMSE = 1.95; MPR RMSE = 2.15). However, we use all four methods for the empirical data sets to check for consistency.

Discussion

Model Assumptions and Robustness

While previous studies have used multi-locus population genetic data to reconstruct the demography and geography of speciation [51-55], here we use single locus mtDNA data to look at patterns across multiple co-distributed taxa. Although single locus inference can be hazardous in the face of coalescent variance and the possibility of selection, our approach offers the possibility to look at patterns of community assembly when the community consists of many non-model organisms where only "barcode" DNA sequence data can be feasibly collected. Not only does our model incorporate the stochasticity of single-locus coalescent variance across taxa, by combining datasets into a hierarchical Bayesian analysis we gain statistical "borrowing strength" [38]. The "borrowing strength" of HABC is achieved by making inferences across groups (i.e. co-distributed taxa) by pooling information across the groups without assuming the groups are from the same population [39,40]. This allows estimating congruence across groups in sub-parameters while borrowing strength from the full comparative phylogeographic sample. This "borrowing strength" translates into higher sample size depending on the magnitude of the "pooling factor" which represents the degree to which sub-parameter estimates (Φ) are pooled together from hyper-parameter estimates of ϕ, rather than estimated independently from each phylogeographic dataset [56].

The possibility of selection at the tightly linked mtDNA genome could bias results of our analytical method [57], especially if balancing selection occurred in the mtDNA genome in ancestral or descendent taxa such that coalescent events were much older than neutral expectations. Likewise, if positive selective sweeps occurred on the mtDNA genome after colonization or vicariance, estimates of Z and $E((\theta_\tau)_2)$ could be biased to reflect colonization [58,59]. However, if positive selection only occurred at the mtDNA genome before colonization or vicariance, then the timing of these isolation events could be better estimated due to reduction of ancestral polymorphism. Furthermore, because selection and demography are ultimately confounding, our method will be less reliable if mitochondrial positive selection is prevalent in the comparative phylogeographic sample. This could be especially troublesome if peripatric speciation by colonization or vicariance involves positive selection at mtDNA genes that allow adaptive divergence in novel peripheral habitats [60-62]. Nonetheless, results from our HABC method can be considered conservative with respect to inferring the geographic and demographic history of isolation across a peripheral community. For example, a strong inference of colonization across an entire data set could result from both strong positive selection and/or small effective colonizing population sizes, but it would be extraordinary if strong positive selection occurred at the mtDNA genome of all Y co-distributed taxon-pairs.

A strong result of temporal concordance in isolation is also conservative with respect to violations from a uniform molecular clock model. If rate variation occurred across the Y taxa, then we would expect an inference in temporal discordance unless rates were inversely proportional with actual isolation times. Nevertheless, this HABC method will be most useful when applied to co-distributed taxon-pairs that are closely related. Our empirical application was restricted to cowrie gastropods (Cypraeidae) and the COI loci we used only marginally rejected rate constancy [63].

Although Bayesian methods are less robust if results are heavily dependent on prior bounds, results from the empirical data were not sensitive to our exploration of different prior assumptions. The overall inferences and hyper-posterior estimates from the empirical data were not sensitive to model assumptions regarding the priors of θ (Additional file 3) or post-isolation migration. Additionally, all four methods of post-acceptance transformation (LLR, PLR, CLR and RAW) yielded identical mode estimates of Z given the Marquesas data and similar mode estimates of Z given the Hawaii data (ranging from 4 – 5).

Another consideration is how deviations from a panmictic Wright-Fisher model could have affected our HABC

estimates. Although the sampling scale is large in some of the source species (Additional file 1), the cowrie gastropod taxa we included have high dispersal capabilities and are therefore not likely to have elevated within species subdivision [63]. This is confirmed by the relatively low levels of within species average pair-wise differences (π_1 and π_2) in both data sets (Additional file 1). If intra-species migration rates are > 1, our idealized coalescent model assuming intra-species panmixia is somewhat appropriate (Slatkin 1985). Even with some population some structure, a standard coalescent model can suffice if a species consists of many small demes with at least moderate migration such that number of demes is approximated by a scaled effective population size [64-67]. If ancestral population structure is approximately scaled by ancestral effective population size, then our chosen priors are conservative because we allow for ancestral population sizes that are two to four times as large as the observed pair-wise distances of extant population sizes (π_1 and π_2; Additional file 1). However, our method should not be applied to populations that are heavily structured over large geographic scales.

Vicariance and Dispersal in Marine Communities

Vicariance and dispersal speciation could be hugely relevant in the marine realm, especially within the highly diverse Indo-Pacific region that is dominated by islands rather than long continuous coastlines. Explanations for elevated Indo-Pacific diversity in the centrally located "coral triangle" portion of this Indo-Pacific region (Philippines, Malay Peninsula, and New Guinea) usually revolve around sympatric speciation followed by outward range shifts [68] or peripatric speciation followed by inward range shifts [69,70]. However, the plausibility of the first hypothesis of sympatric speciation is very controversial on theoretical grounds [71,72] as well as being very difficult to test empirically [18]. On the other hand, the second hypothesis of peripatric speciation is a much more likely force behind Indo-Pacific diversification if long distance oceanic dispersal to peripheral populations is sufficiently low for isolation and subsequent reproductive or ecological divergence to emerge between central and peripheral archipelagoes [1-7].

Instead of the classic vicariance model, under our marine vicariance model (or "soft vicariance") an ancestral range is inter-connected by long distance gene flow that is interrupted by oceanographic changes in temperature, sea level and/or currents [10-12]. In this case we might predict that co-distributed peripheral endemics became isolated simultaneously in taxa with lower dispersal capability. On the other hand, our marine colonization model is more similar to the classic dispersal or peripatric model. Here, allopatric isolation arises via sweepstakes colonization where the timing of colonization could be predicted to occur randomly across the co-distributed endemics after an archipelago emerges from geological processes.

Hawaiian vs Marquesas Dynamics

Both Hawaii and the Marquesas have some of the highest levels of endemism in all of Oceania [73-75], but HABC analyses support different histories for their endemic species. The Hawaiian archipelago is the most isolated island chain in the Indo-West Pacific and has existed for > 70 My with coral reefs since at least 35 My [76]. It is now well established that terrestrial endemics can be older than the oldest emergent island in the archipelago, a pattern resulting from initial colonization of older, now subsided islands, followed by dispersal to new islands after emergence [77]. Thus, there has been ample opportunity for isolation and speciation, perhaps even more so for marine taxa. Although the lower sample sizes lead to greater uncertainty associated with the Hawaiian estimates (Figure 3), the hyper-parameter 95% credibility intervals suggest a strong inference of isolation via soft vicariance in at least a subset of the Hawaiian taxon-pairs (Z = 0.0 – 9.22; $E((\theta_\tau)_2)$ = 0.30–0.95). If this is the case, then occupation of the Hawaiian archipelago was much older than the Marquesas archipelago, consistent with geologic evidence. Moreover, the inference of soft vicariance in a number of the taxon pairs suggests that there was greater potential for migrants between this archipelago and the central Pacific and Indo-Pacific triangle regions during older periods. Such connectivity of the Hawaiian chain to the remaining Indo-West Pacific via Johnston Atoll has been suggested in other studies [78-80].

In contrast, we find strong inference of isolation via colonization across all seven cowrie gastropod endemics co-distributed in the Marquesas, as well as a strong inference of temporal concordance in this colonization. Unlike other island chains in the Pacific Ocean that have older seamounts and atolls trending away from most recent island (e.g. Hawaii), the Marquesan hotspot is unusual in being quite young. The ages of Marquesan islands range from 1.3 My (Fatu Hiva) to 6.0 My (Eiao) [81]. If we apply a molecular clock of 1% divergence per My [63], the inferred timing of simultaneous colonization of the Marquesas archipelago is from 0.84 – 1.90 My (Additional file 3A), and is consistent with the young origins of the islands. If we accept our strong inference of temporal concordance, it could be argued that this assemblage of cowries colonized via an episodic oceanographic event that caused a surge in gene flow from the central Pacific region. Given that the Marquesas has one of the highest levels of marine endemism in Oceania [73,74], it will be interesting to see if HABC analyses on other taxon-pairs show similarly young divergences and temporal congruence.

Conclusion

Although soft vicariance and colonization are likely to produce relatively similar genetic patterns when only a single taxon-pair is considered, our simulation analysis shows that our hierarchical Bayesian model can potentially detect if either history is a dominant process across a marine community. The empirical implementation of our method yields a strong inference of isolation via simultaneous colonization across all seven cowrie gastropod endemics co-distributed in the Marquesas. In contrast, our method shows a strong inference of isolation via soft vicariance in at least a subset of the 11 Hawaiian taxon-pairs, although the smaller sample size resulted in less certainty in our estimates.

Our HABC method exemplifies the utility in "statistical phylogeographic" approaches [82,83] rather then qualitative and descriptive approaches that make large inferences from the small details observed from gene trees [84]. The HABC approach accomplishes this by analyzing all the phylogeographic datasets at once in order to make across taxon-pair inferences about biogeographic processes while explicitly allowing for uncertainty in the demographic differences within each taxon-pair.

Although the approach described here uses HABC to test for two particular biogeographic explanations of allopatric diversification across co-distributed taxa, the HABC framework is flexible and therefore can provide a skeleton for testing other biogeographic models from comparative phylogeographic data. Indeed, one of the original objectives of phylogeography comparative was to resolve deep-seated questions about how climate change drives community assembly and evolution of whole biotas [85]. However, this goal has so far been unrealized [86-88] because comparative phylogeographic studies rarely involve more than a handful of co-distributed species. Comparative phylogeographic datasets are bound to have explosive growth as collecting DNA sequence data across a wide diversity of co-distributed taxa scales up to the level of comprehensive ecosystem sampling. Such "community-scale" comparative phylogeographic data sets could potentially test classic biogeographic hypotheses (e.g. vicariance versus dispersal) at the community level [89,90], as well as test controversial and fundamental hypotheses in community ecology such as Hubbell's Neutral theory [91], Tillman's stochastic competitive assembly model [92], and Diamond's niche assembly rules [93,94]. As comparative phylogeographic datasets grow to include > 100 co-distributed taxon-pairs, the HABC approach will be well suited to dissect temporal patterns in community assembly and thereby provide a bridge linking comparative phylogeography with community ecology.

Abbreviations

HABC: Hierarchical Approximate Bayesian Computation; mtDNA: Mitochondrial DNA; COI: Cytochrome oxidase 1.

Authors' contributions

MH developed, tested, and implemented the HABC model. CM collected and sequenced the cowrie data and provided feedback on model development. Both MH and CM contributed to the writing.

Additional material

Additional file 1

Table 1 – Ranges, samples sizes and three summary statistics of cowrie data. *(A) Seven Marquesas cowrie sister taxon pairs (Cypraeidae) (B) Eleven Hawaiian cowrie sister taxon pairs.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-322-S1.doc]

Additional file 2

Table 2 – Summary of hierarchical model parameters. *Summary of the hierarchical model. Times are referred to as times before the present. (A) Sample size configuration (B) Hyper-parameters (C) Sub-parameters (D) Hyper-parameter summaries. * Estimated under the stage 1 general model. **Only estimated under a constrained stage 2 model (Z constrained to be the integer closest to the posterior mode estimate generated from the stage 1 analysis).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-322-S2.doc]

Additional file 3

Table 3 – Posterior mode hyper-parameter estimates and their 95% credibility intervals. *Hyper-posterior mode estimates and their 95% credibility intervals from two comparative phylogeographic data sets of taxon-pairs that include endemic species or subspecies in the (A & B) Marquesas and (C & D) Hawaiian archipelagoes. Average colonization $E(\tau_C)$ and vicariance $E(\tau_V)$ times are given in units of My by assuming a divergence rate of 1% per My. Bold values are obtained from stage 1 of the analysis, whereas the remaining estimates are obtained from the stage 2 analysis where Z and $E((\theta_\tau)_2)$ are held to their estimated values obtained in stage 1.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-322-S3.doc]

Acknowledgements

We thank M. Beaumont for kindly providing R scripts used in the HABC algorithm; W. Huang, E. Stahl and N. Takebayashi for collaboration in the ongoing development of the MSBAYES software pipeline; and the International Biogeography Society for inviting us to participate in the special symposium in marine biogeography at the 2007 meeting. We thank A. Pyron for coining the term "soft vicariance". M. Beals, W. Woodman and D. Watts are gratefully acknowledged for providing key specimens. M. Hickerson was supported by NSF (DEB 073648) and C. Meyer was supported by NSF (DEB-9807316 & 0316338, OCE-0221382). We also greatly thank the

anonymous reviewers and the communicating editor H. Zauner for useful comments and suggestions to improve the manuscript.

References

1. Mayr E: **Geographic speciation in tropical echinoids.** *Evolution* 1954, **8:**1-18.
2. Mayr E: **Animal species and evolution.** Cambridge, MA: Harvard University Press; 1963.
3. Zigler KS, McCartney MA, Levitan DR, Lessios HA: **Sea urchin bindin divergence predicts gamete compatibility.** *Evolution* 2005, **59:**2399-2404.
4. Palumbi SR: **Genetic divergence, reproductive isolation, and marine speciation.** *Annu Rev Ecol Syst* 1994, **25:**547-572.
5. Palumbi SR, Lessios HA: **Evolutionary animation: How do molecular phylogenies compare to Mayr's reconstruction of speciation patterns in the sea?** *Proc Natl Acad Sci USA* 2005, **102:**6566-6572.
6. Palumbi S: **Marine speciation on a small planet.** *TREE* 1992, **7(4):**114-118.
7. Paulay G, Meyer C: **Dispersal and divergence across the greatest ocean region: do larvae matter?** *Integr Comp Biol* 2006, **46:**269-281.
8. **Briggs: Centrifugal speciation and centres of origin.** *Journal of Biogeography* 2008, **27:**1183-1188.
9. Briggs JC: **The marine East Indies: diversity and speciation.** *Journal of Biogeography* 2005, **32:**1517-1522.
10. Paulay G, Meyer C: **Diversification in the Tropical Pacific: Comparisons Between Marine and Terrestrial Systems and the Importance of Founder Speciation.** *Integr Comp Biol* 2002, **42:**922-934.
11. Barber PH, Palumbi SR, Erdmann MV, Moosa MK: **Sharp genetic breaks among populations of *Haptosquilla pulchella* (Stomatopoda) indicate limits to larval transport: patterns, causes, and consequences.** *Mol Ecol* 2002, **11(4):**659-674.
12. Barber PH, Palumbi SR, Erdmann MV: **Comparative phylogeography of three co-distributed stomatopods: origins and timing of regional lineage diversification in the coral triangle.** *Evolution* 2006, **60:**1825-1839.
13. Ladd HS: **Origin of the Pacific Island molluscan fauna.** *American Journal of Science* 1960, **258A:**137-150.
14. Jokiel P, Martinelli FJ: **The vortex model of coral reef biogeography.** *Journal of Biogeography* 1992, **19:**449-458.
15. Waters JM, Dijkstra LH, Wallis GP: **Biogeography of a southern hemisphere freshwater fish: how important is marine dispersal?** *Mol Ecol* 2000, **9:**1815-1821.
16. Yoder AD, Nowak MD: **Has Vicariance or Dispersal Been the Predominant Biogeographic Force in Madagascar? Only Time Will Tell.** *Annual Review of Ecology, Evolution, and Systematics* 2006, **37:**405-431.
17. Chesser RT, Zink RM: **Modes of Speciation in Birds: A Test of Lynch's Method.** *Evolution* 1994, **48:**490-497.
18. Losos JB, Glor RE: **Phylogenetic comparative methods and the geography of speciation.** *Trends Ecol Evol* 2003, **18:**220-227.
19. Wiley EO: **Vicariance Biogeography.** *Annu Rev Ecol Syst* 1988, **19:**513-542.
20. Wiley EO: **Phylogenetic Systematics and Vicariance Biogeography.** *Systematic Botany* 1980, **5:**194-220.
21. Nelson G, Platnick NI: **Systematics and biogeography: cladistics and vicariance.** New York: Columbia University Press; 1981.
22. Brooks DR: **Hennigs Parasitological method, a proposed solution.** *Syst Zool* 1981, **30(3):**229-249.
23. Ronquist F: **Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography.** *Syst Biol* 1997, **46:**195-203.
24. Zink RM, Blackwell-Rago RC, Ronquist F: **The shifting roles of dispersal and vicariance in biogeography.** *Proceedings of the Royal Society of London, Series B* 2000, **267:**497-503.
25. Ree RH, Moore BR, Webb CO, Donoghue MJ: **A likelihood framework for inferring the evolution of geographic range on phylogenetic trees.** *Evolution* 2005, **59(11):**2299-2311.
26. Ree RH, Smith SA: **Maximum-likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis.** *Syst Biol* 2008, **57:**4-14.
27. Sanmartín I, Mark P van der, Ronquist F: **Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands.** *Journal of Biogeography* 2007, **35:**428-449.
28. Noonan BP, Chippindale PT: **Vicariant origin of Malagasy reptiles supports Late Cretaceous Antarctic landbridge.** *Am Nat* 2006, **168:**730-741.
29. de Queiroz K: **The resurrection of oceanic dispersal in historical biogeography.** *Trends Ecol Evol* 2005, **20:**68-73.
30. Waters JM: **Driven by the West Wind Drift? A synthesis of southern temperate marine biogeography, with new directions for dispersalism.** *Journal of Biogeography* 2008, **35:**417-427.
31. Cunningham CW, Collins TM: **Beyond area relationships: Extinction and recolonization in molecular marine biogeography.** In *Molecular Ecology and Evolution: Approaches and Applications* 2nd edition. Edited by: Schierwater B, Streit B, Wagner G, DeSalle R. Basel, Switzerland: Birkhauser Verlag; 1998:297-322.
32. Waters JM, Craw D: **Goodbye Gondwana? New Zealand Biogeography, Geology, and the Problem of Circularity.** *Syst Biol* 2006, **55:**351-356.
33. Brumfield RT, Edwards SV: **Evolution into and out of the Andes: a Bayesian analysis of historical diversification in Thamnophilus antshrikes.** *Evolution* 2007, **61(2):**346-367.
34. Harrison RG: **Molecular changes at speciation.** *Annu Rev Ecol Syst* 1991, **22:**281-308.
35. Hickerson MJ, Stahl E, Lessios HA: **Test for simultaneous divergence using approximate Bayesian computation.** *Evolution* 2006, **60:**2435-2453.
36. Hickerson MJ, Stahl E, Takebayashi N: **msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation.** *BMC Bioinformatics* 2007, **8:**268.
37. Slatkin M: **Gene flow in natural populations.** *Ann Rev Ecol Syst* 1985, **16:**393-430.
38. Gelman A, Carlin JB, Stern HS, Rubin DB: **Bayesian Data Analysis.** London, UK: Chapman and Hall/CRC; 1995.
39. James W, Stein C: **Estimation with quadradic loss.** In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: 1960* Berkeley, CA: University of California Press; 1960.
40. Beaumont MA, Rannala B: **The Bayesian revolution in genetics.** *Nat Rev Genet* 2004, **5:**251-261.
41. Leaché A, Crews SA, Hickerson MJ: **Did an ancient seaway across Baja California cause community isolation?** *Biology Letters* 2007, **3:**646-650.
42. Beaumont MA, Zhang W, Balding DJ: **Approximate Bayesian computation in population genetics.** *Genetics* 2002, **162:**2025-2035.
43. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585-595.
44. Gelman A, Carlin JB, Stern HS, Rubin DB: **Bayesian Data Analysis.** Boca Raton, FL: Chapman and Hall/CRC; 2004.
45. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation.** *Bioinformatics* 2002, **18:**337-338.
46. Tajima F: **Evolutionary relationship of DNA sequences in finite populations.** *Genetics* 1983, **105:**437-460.
47. Kass RE, Raftery A: **Bayesian factors.** *Journal of the American Statistical Association* 1995, **90:**773-795.
48. Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L: **Statistical evaluation of alternative models of human evolution.** *Proc Natl Acad Sci USA* 2007, **104:**17614-17619.
49. François O, Blum MGB, Jakobsson M, Rosenberg NA: **Demographic History of European Populations of Arabidopsis thaliana.** *PLoS Genetics* 2008, **4:**e1000075.
50. Beaumont MA: **Joint determination of topology, divergence time and immigration in population trees.** In *Simulations, Genetics and Human Prehistory* Edited by: Matsumura S, Forster P, Renfrew C. Cambridge: McDonald Institute for Archaeological Research; 2008:135-154.
51. Rosenblum E, Hickerson MJ, Moritz C: **A multilocus perspective on colonization accompanied by selection and gene flow.** *Evolution* 2007, **61:**2971-2985.
52. Machado CA, Kilman RM, Market J, Hey J: **Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives.** *Mol Biol Evol* 2002, **19:**472-488.
53. Hey J, Nielsen R: **Multilocus methods for estimating population sizes, migration rates and divergence time, with applications**

to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 2004, **167:**747-760.
54. Osada N, Wu C-I: **Inferring the mode of speciation with genomic data – Examples from the great apes.** *Genetics* 2005, **169:**259-264.
55. Zhou R, Zeng K, Wu W, Chen X, Yang Z, Shi S, Wu C-I: **Population Genetics of Speciation in Nonmodel Organisms: I. Ancestral Polymorphism in Mangroves.** *Mol Biol Evol* 2007, **24:**2746-2754.
56. Gelman A, Hill J: **Data analysis using regression and multilevel/hierarchical models.** New York, NY: Cambridge University Press; 2007.
57. Hahn MW: **Towards a selection theory of molecular evolution.** *Evolution* 2008, **62:**255-265.
58. Gillespie JH: **Is the population size of a species relevant to its evolution.** *Evolution* 2001, **55:**2161-2169.
59. **Bazin: Population Size Does Not Influence Mitochondrial Genetic Diversity in Animal.** *Science* 2006, **312:**570-572.
60. Hendry AP, Nosil P, Rieseberg LH: **The speed of ecological speciation.** *Functional Ecology* 2007, **21:**455-464.
61. Gavrilets S, Vose A: **Dynamic patterns of adaptive radiation.** *Proc Natl Acad Sci USA* 2005, **13:**18040-18045.
62. Nosil P, Vines TH, Funk DJ: **Reproductive isolation caused by natural selection against immigrants from divergent habitats.** *Evolution* 2005, **59:**705-719.
63. Meyer CP, Paulay G: **DNA barcoding: Error rates based on comprehensive sampling.** *PLoS Biol* 2005, **3(12):**e422.
64. Wakeley J, Lessard S: **Corridors for migration between large subdivided populations, and the structured coalescent.** *Theor Popul Biol* 2006, **70:**412-420.
65. Wakeley J, Takahashi T: **The many-demes limit for selection and drift in a subdivided population.** *Theor Popul Biol* 2004, **66:**83-91.
66. Wakeley J: **Recent trends in population genetics: More data! More math! Simple models?** *Journal of Heredity* 2004, **95:**397-405.
67. Notohara M: **The strong migration limit for the genealogical process in geographically structured populations.** *J Math Biol* 1997, **36(2):**188-200.
68. Briggs JC: **Marine centres of origin as evolutionary engines.** *Journal of Biogeography* 2003, **30:**1-18.
69. Briggs JC: **The marine East Indies: center of origin?** *Global Ecology and Biogeography* 1992, **2:**149-156.
70. Briggs JC: **Modes of speciation: Marine Indo-West Pacific.** *Bull Mar Sci* 1999, **65:**615-656.
71. Coyne JA, Orr HA: **Speciation.** Sunderland, MA: Sinauer Associates Inc; 2004.
72. Gavrilets S: **Fitness landscapes and the origin of species.** Princeton, NJ: Princeton University Press; 2004.
73. Rehder HA: **The Marine Molluscan Fauna of the Marquesas Islands.** *American Malacological Union* 1968:29-32.
74. Randall JE: **Zoogeography of shore fishes of the Indo-Pacific region.** *Zool Stud* 1998, **37:**227-268.
75. Kay EA, Palumbi SR: **Endemism and evolution in Hawaiian, USA marine invertebrates.** *Trends Ecol Evol* 1987, **2(7):**183-186.
76. Grigg RW: **Paleoceanography of coral reefs in the Hawaiian-Emperor Chain – revisted.** *Coral Reefs* 1997, **16:**S33-S38.
77. Fleischer RC, McIntosh CE, Tarr CL: **using phylogeographic reconstructions and K-Ar-based ages of the Hawaiian Islands to estimate molecular evolutionary rates.** *Mol Ecol* 1998, **7(4):**533-545.
78. Kobayashi DR: **Colonization of the Hawaiian Archipelago via Johnston Atoll: a characterization of oceanic transport corridors for pelagic larvae using computer simulation.** *Coral Reefs* 2006, **25:**407-417.
79. Rivera MJ, Kelley CD, Roderick GK: **Subtle population genetic structure I nthe Hawaiian grouper, Epinephelus quernus (Serranidae) as revealed by mitochondrial DNA analyses.** *Biol J Linn Soc* 2004, **81:**449-468.
80. Grigg RW: **Acropora in Hawaii USA 2. Zoogeography.** *Pacific Science* 1981, **35:**1-13.
81. McNutt M, Bonneville A: **A shallow, chemical origin for the Marquesas Swell.** *Geochemistry Geophysics Geosystems* 2000, **1:**.
82. Smouse PE: **To tree or not to tree.** *Mol Ecol* 1998, **7:**399-412.
83. Knowles LL, Maddison WP: **Statistical phylogeography.** *Mol Ecol* 2002, **11(12):**2623-2635.
84. Panchel M, Beaumont BA: **The automation and evaluation of nested clade phylogeographic analysis.** *Evolution* 2007, **61(6):**1466-1480.
85. Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC: **Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics.** *Ann Rev Ecol Syst* 1987, **18:**489-522.
86. Avise JC: **Phylogeography: The history and formation of species.** Cambridge: Harvard University Press; 2000.
87. Bermingham E, Moritz C: **Comparative phylogeography: concepts and applications.** *Mol Ecol* 1998, **7:**367-369.
88. Arbogast BS, Kenagy GJ: **Comparative phylogeography as an integrative approach to historical biogeography.** *Journal of Biogeography* 2001, **28:**819-825.
89. Rosen DE: **Vicariant Patterns and Historical Explanation in Biogeography.** *Systematic Zoology* 1978, **27:**159-188.
90. Carlquist C: **The biota of long-distance dispersal, I: Principles of dispersal and evolution.** *Quarterly Review of Biology* 1966, **41:**247-270.
91. Hubbell SP: **The Unified Neutral Theory of Biodiversity and Biogeography.** Princeton, NJ: Princeton University Press; 2001.
92. Tillman D: **Niche tradeoffs, neutrality, and community structure: A stochastic theory of resource competition, invasion, and community assembly.** *Proc Natl Acad Sci USA* 2004, **101:**10854-10861.
93. Gotelli NJ, McCabe DJ: **Species co-occurrence: a meta-analysis of j. M. Diamond's assembly rules model.** *Ecology* 2002, **83:**2091-2096.
94. Diamond JM: **Assembly of Species Communities.** In *Ecology and Evolution of Communities* Edited by: Cody ML, Diamond JM. Belknap: Harvard; 1975:342-444.