MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 1345–1359 doi: 10.1111/j.1365-294X.2012.05475.x

Influence of finite-sites mutation, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity

ASHER D. CUTTER,*†[1] GUO-XIU WANG,*‡[1] HUI AI‡ and YUE PENG‡

*Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON M5S 3B2, Canada, †Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada, ‡Hubei Key Laboratory of Genetic Regulation and Integrative Biology, HuaZhong Normal University, Wuhan, Hubei, China

Abstract

Molecular hyperdiversity has been documented in viruses, prokaryotes and eukaryotes. Such organisms undermine the assumptions of the infinite-sites mutational model, because multiple mutational events at a site comprise a non-negligible portion of polymorphisms. Moreover, different sampling schemes of individuals from species with subdivided populations can profoundly influence resulting patterns and interpretations of molecular variation. Inspired by molecular hyperdiversity in the nematode *Caenorhabditis* sp. 5, which exhibits average pairwise differences among synonymous sites of >5% as well as modest population structure, we investigated via coalescent simulation the joint effects of a finite-sites mutation (FSM) process and population subdivision on the variant frequency spectrum. From many demes interconnected through a stepping-stone migration model, we constructed local samples from a single deme, pooled samples from several demes and scattered samples of a single individual from numerous demes. Compared with a single panmictic population at equilibrium, we find that high population mutation rates induce a deficit of rare variants (positive Tajima's *D*) under a FSM model. Population structure also induces such a skew for local samples when migration is high and for pooled samples when migration is low. Contrasts of sampling schemes for *C.* sp. 5 imply high mutational input coupled with high migration. We propose that joint analysis of local, pooled and scattered samples for species with subdivided populations provides a means of improving inference of demographic history, by virtue of the partially distinct patterns of polymorphism that manifest when sequences are analyzed according to differing sampling schemes.

Keywords: molecular evolution, phylogeography, population genetics—empirical, population genetics—theoretical, worms, *Caenorhabditis*

Received 15 October 2011; revision received 12 December 2011; accepted 14 December 2011

Introduction

Emerging examples of molecular hyperdiversity in eukaryotes, in which population polymorphism at neutral nucleotide sites exceeds 5%, indicate that the consequences of high population mutation rates need more widespread consideration beyond microbes (McVean *et al.* 2002; Perez-Losada *et al.* 2006; Desai & Plotkin 2008). A crucial issue for population genetic analysis for such organisms is the invalidated assumption of the infinite-sites mutation model (ISM) that a given site will be hit by mutation once at most (Nei & Li 1979). The ISM is a cornerstone of theory in molecular population genetics because of its simplicity and relevance to the many organisms that have relatively low levels of polymorphism. However, finite-sites mutation models (FSM) become more appropriate as the likelihood of multiple mutational hits at a site becomes non-negligible

Correspondence: Asher D. Cutter, Fax: 416 978 5878;
E-mail: asher.cutter@utoronto.ca
[1]These authors contributed equally to this work.

© 2012 Blackwell Publishing Ltd

[roughly, when the population mutation parameter $\theta \geq 5\%$ per site (Desai & Plotkin 2008)], such as for species of obligatorily outbreeding *Caenorhabditis* nematodes like *C.* sp. 5 (Wang *et al.* 2010).

Several early studies of FSMs focused on among-site rate heterogeneity, exploring its influence on the site frequency spectrum of polymorphic nucleotide variants within populations (Bertorelle & Slatkin 1995; Aris-Brosou & Excoffier 1996; Rogers *et al.* 1996; Yang 1996). Because tests of neutrality that depend on the population frequency distribution of variants across sites can be powerful indicators of demographic history, as well as of perturbations by natural selection, it is important to understand how FSMs affect the site frequency spectrum. Among-site mutation rate variability leads to positive skews in the site frequency spectrum summary metric, Tajima's D (D_{Taj}) (Tajima 1989). In other words, relative to the expectation that $D_{\mathrm{Taj}} \cong 0$ under the standard neutral model for a single panmictic population at equilibrium with an ISM, high mutation rates under an FSM appear to yield a deficit of low-frequency variants. This happens because highly mutable sites receive multiple mutations that can increase the number of pairwise differences between sequences (π) without increasing the number of segregating sites (S); D_{Taj} is proportional to the difference between π and S. Even with a uniform mutation rate among sites, a similar effect should result when the population mutation rate is sufficiently high. Of course, the extent to which this is evident might be counteracted to some degree if back-mutations occur at sites carrying common derived variants. Ordinarily, the finding of $D_{\mathrm{Taj}} > 0$ across many loci would be interpreted as evidence of non-equilibrium population dynamics, such as population contraction or pooling of samples from multiple subpopulations (Tajima 1993; Stadler *et al.* 2009). Recently, Jenkins & Song (2011) derived a model for the site frequency spectrum under a tri-allelic FSM. Alternative FSMs using a modified Poisson random field approach have been developed that permit appropriate inference about selection from the spectrum of site frequencies in hyperdiverse populations (Desai & Plotkin 2008). McVean *et al.* (2002) also developed approximate FSM estimators of the population mutation rate (θ_{w}^{*}) and recombination rate. However, it is unclear how these FSM approaches compare under complex demographic scenarios.

Population demographic complexity can occur along many axes, with one common feature being population structure. The geographic ranges for many species are subdivided into local demes, and this population subdivision can modify levels and patterns of polymorphism in important ways. In particular, sampling scheme can strongly affect population genetic inferences (Ptak & Przeworski 2002; Stadler *et al.* 2009). For example, if a sample of multiple individuals deriving from two or more demes are 'pooled' together, then diversity estimates will be inflated and the distribution of variant frequencies will tend to be skewed towards an excess of intermediate-frequency variants, relative to the expectation from a standard neutral model for a single panmictic population (Stadler *et al.* 2009). An important means of conceptualizing the effects of structure in a coalescent framework is the notion of separation of timescales (Wakeley 1999), such that coalescence within a deme can be thought to occur more rapidly than among demes, and these two phases can be considered separately. Consequently, 'scattered samples' comprised of a single representative from each of many demes will recapitulate the neutral coalescent process, unlike for 'pooled' samples (Wakeley 1999; Lessard & Wakeley 2004). Importantly, patterns of polymorphism for a 'local sample' from a single deme embedded within a large matrix of subpopulations also can deviate substantially from that expected for a single population in isolation (Pannell 2003; Ray *et al.* 2003; De & Durrett 2007; Stadler *et al.* 2009). This effect is stronger for stepping-stone models of migration where demes share migrants only in a local neighbourhood than it is for island migration models in which demes share migrants regardless of their location in the network of subpopulations (De & Durrett 2007; Stadler *et al.* 2009). While pooled samples show greatest variant frequency skew with low migration among demes, local samples show greatest skew at intermediate rates of migration (Stadler *et al.* 2009). Exploration of sampling schemes in this vein for empirical nucleotide polymorphism data has been conducted in several species, most notably a recent analysis of wild tomatoes (Tellier *et al.* 2011), although contrasts of sampling schemes relative to null expectations have not yet been exploited to their full potential (reviewed in Stadler *et al.* 2009).

These explicit studies of sampling scheme in structured populations followed an ISM. Chikhi *et al.* (2010) considered a similar approach with a stepwise-mutation model (SMM), appropriate for microsatellite loci. While the SMM incorporates repeated mutation at the same locus as do FSMs, it is otherwise not applicable to single-nucleotide evolution; nevertheless, high SMM mutation rates exacerbate spurious demographic signals in samples from structured populations (Chikhi *et al.* 2010). Thus, there lies an important gap in understanding how sampling schemes from structured populations might interact with FSM to contribute to deviations from standard neutral expectations about patterns of polymorphism.

Caenorhabditis sp. 5 is a nematode relative of *Caenorhabditis elegans* that is found throughout, but apparently restricted to, East Asia. This species exhibits both very

© 2012 Blackwell Publishing Ltd

high population polymorphism, as well as genetic isolation by distance, implicating some degree of population structure across its range (Wang *et al.* 2010). Consequently, it is crucial to determine the interplay between FSM and different sampling schemes from structured populations under well-defined conditions to derive appropriate inferences about population history for organisms like *C.* sp. 5. Here, we use new population polymorphism data for *C.* sp. 5 as an empirical reference and motivator for coalescent simulations that explore generally the patterns of polymorphism that occur under an FSM with structured populations. We find that high population mutation rates do indeed exacerbate the effect of population structure on diversity estimates and the site frequency spectrum. We argue that more informed inferences about the demographic history of populations can be obtained by contrasting patterns of polymorphism for different sampling schemes.

Materials and methods

Strain sampling and molecular methods

We isolated 19 new isofemale lines (strains) of *C.* sp. 5 from five localities across three provinces in China (Table 1), using standard collection procedures (Barrière & Félix 2006). A map of *C.* sp. 5 strains is available online at http://www.google.com/maps/ms?msa=0&msid=117700919974655793194.00046c90c5fa3a4c5837b&t=h&z=2. We amplified by PCR and sequenced 10 X-linked loci for these strains, using seven sets of previously published primers and three new loci (Cutter 2008; Wang *et al.* 2010). We combined these new data with a previously published data set for 22 previously described strains for seven loci (Wang *et al.* 2010). In addition, we amplified and sequenced three new X-linked nuclear coding loci and one new mitochondrial coding locus for the combined set of 41 strains; each locus represents a partial coding sequence of a gene. The full data set includes a total of 7981 bp per strain, or 6283 bp for the 41 strains excluding all gaps and missing data with 1388 synonymous sites. The primers for these four new loci are *Csp5-dgn-1* (forward AGGCTGGACAACTCGCTTC, reverse CCCTTG GAGACATATTCGGT or TTGCCGCTCTTCTTGATGC), *Csp5-T21B6.3* (forward TTGACCAGCATGCTCAACTC, reverse TTGGCTTTCCTCATTTGGTC), *Csp5-D1005.1* (forward GAATCAAGCCAGGAGCATTC, reverse GGA ACCACAACAATCCCAAA) and *Csp5-COII* (forward TTCAGTATGTTTGACTTCCA, reverse GCTCTATATTC TCTTACACCACA from (Thomas & Wilson 1991)). We isolated and amplified genomic DNA from single male animals using the Qiagen Repli-g mini kit as described (Cutter 2008), so sequences from the X-linked loci represent a single haplotype that avoids complications of heterozygous base-calling and phasing of sequence traces. Sequencing of both orientations of PCR amplicons was performed at the University of Arizona UAGC sequencing facility. New sequences have been deposited in Genbank under accessions JQ219362–JQ219658.

Molecular population genetic analysis

We computed population genetic statistics for the *C.* sp. 5 samples using several sampling schemes. First, we pooled all 41 strains from the 13 geographic localities (or, 25 strains from five localities pooled). Second, we pooled 34 of the strains by province of collection for the three provinces with >10 strains (Guangdong $n = 11$, Hubei $n = 11$, Sichuan $n = 12$). Third, we treated each sampling locality separately, restricting analysis to the five localities with four or more strains. Finally, we constructed 1000 replicate 'scattered samples' by randomly including one strain from each of the 13 geographic localities in our data set. Such scattered samples in a structured population should approximate the standard neutral coalescent process (Wakeley 1999; Lessard & Wakeley 2004). In all cases, we computed measures of molecular variation (Jukes-Cantor adjusted average number of pairwise differences per synonymous site, $\pi_{\text{syn-JC}}$) and the variant frequency spectrum [Tajima's D_{Taj} and Schaeffer's $D_{\text{Taj/min}}$ = Tajima's D/D_{min} for synonymous sites, (Schaeffer 2002)] for each locus separately with Polymorphorama (Bachtrog & Andolfatto 2006). Polymorphorama calculates D_{Taj} using the estimated number of mutations, rather than the observed number of segregating sites; for some analyses, DNASP 5.10 was used to calculate D_{Taj} using the observed number of segregating sites (Librado & Rozas 2009). Multipopulation analyses used DNASP to calculate genetic differentiation (F_{st}, G_{st}, D_{a}, K_{s}^{*} , K_{st}^{*}) across all sites (Librado & Rozas 2009), from which the standardized metric G'_{st} was computed (Hedrick 2005). D_{a} represents the net number of substitutions per site between populations, after accounting for polymorphism within each (Nei & Li 1979); K_{s}^{*} and K_{st}^{*} are sensitive metrics of structure for samples with high diversity (Hudson *et al.* 1992; Hudson 2000). For the scattered samples, we calculated the mean of $\pi_{\text{syn-JC}}$ and D_{Taj} across the 10 nuclear loci and compared the distribution of 1000 means to the observed mean $\pi_{\text{syn-JC}}$ and D_{Taj} for all strains; similarly, we compared the distribution of 1000 mitochondrial scattered samples to the observed values of $\pi_{\text{syn-JC}}$ and D_{Taj}.

Finite-sites mutation model coalescent simulations

To investigate the influence of a finite-sites mutational process on the site frequency spectrum, we simulated

© 2012 Blackwell Publishing Ltd

Table 1 Strains used in this study

Strain	Latitude	Longitude	Locality	Date	Source	References
GXW0002	30.514N	114.495E	Wuhan (Jiufeng village), Hubei	4/4/2010	Soil from vegetable garden	This study
GXW0003	30.514N	114.495E	Wuhan (Jiufeng village), Hubei	4/4/2010	Soil from grove	This study
GXW0004	30.543N	114.420E	Wuhan(Botanical Garden), Hubei	5/11/2010	Soil from parterre	This study
GXW0005	30.543N	114.420E	Wuhan(Botanical Garden), Hubei	5/11/2010	Soil near pond	This study
GXW0006	30.543N	114.420E	Wuhan(Botanical Garden), Hubei	6/2/2010	Soil under chrysanthemum	This study
GXW0007	30.543N	114.420E	Wuhan(Botanical Garden), Hubei	6/2/2010	Soil under Chinese parasol tree	This study
GXW0008	30.676N	104.100E	Chengdu, Sichuan	7/21/2010	Soil under decaying bamboo leaves	This study
GXW0009	29.352N	102.411E	Yaan, Sichuan	9/7/2010	Soil under decaying leaves	This study
GXW0010	29.385N	102.626E	Yaan, Sichuan	9/7/2010	Black soil	This study
GXW0011	29.398N	102.372E	Yaan, Sichuan	9/7/2010	Soil under decaying leaves	This study
GXW0012	29.981N	103.001E	Yaan, Sichuan	9/7/2010	Soil near a apartment	This study
GXW0013	29.981N	103.001E	Yaan, Sichuan	9/7/2010	Soil under Osmanthus tree	This study
GXW0014	30.676N	104.100E	Chengdu, Sichuan	9/7/2010	Compost soil under bamboo	This study
GXW0015	30.676N	104.100E	Chengdu, Sichuan	9/9/2010	Compost soil under bamboo	This study
GXW0016	30.636N	104.090E	Chengdu, Sichuan	9/9/2010	Soil in lawn	This study
GXW0017	30.636N	104.090E	Chengdu, Sichuan	9/9/2010	Soil in lawn	This study
GXW0018	30.636N	104.090E	Chengdu, Sichuan	9/9/2010	Soil in parterre	This study
GXW0019	25.454N	119.008E	Putian, Fujian	8/5/2010	Soil under litchi tree	This study
GXW0020	25.454N	119.008E	Putian, Fujian	8/5/2010	Soil under longan tree	This study
JU1201	31.317N	120.617E	Suzhou, Jiangsu	8/17/2007	Small fruit in garden	Wang *et al.* (2010)
JU1202	30.300N	120.183E	Hangzhou, Zhejiang	8/25/2007	Small fruit in park	Wang *et al.* (2010)
JU1423	22.417N	105.583E	Bac Kan(near Ba Be Lake), Vietnam	4/29/2008	Rotting fruit	Wang *et al.* (2010)
JU727	25.903N	109.651E	Chengyang, Guangxi	5/6/2005	Under a tree	Wang *et al.* (2010)
VX0035	25.028N	113.393E	Lechang (Changlai town), Guangdong	9/1/2008	Compost soil in cabbage field	Wang *et al.* (2010)
VX0036	25.028N	113.393E	Lechang (Changlai town), Guangdong	9/1/2008	Snail (Oxychilus sp.)	Wang *et al.* (2010)
VX0037	25.028N	113.393E	Lechang (Changlai town), Guangdong	9/1/2008	Rotting bamboo leaves (bambusa textilis)	Wang *et al.* (2010)
VX0065	30.514N	114.495E	Wuhan (Jiufeng village), Hubei	9/20/2008	Soil near pond	Wang *et al.* (2010)
VX0066	29.350N	102.683E	Yaan, Sichuan	9/5/2008	Soil under poplar tree	This study
VX0067	30.550N	114.250E	Wuhan (Hanyang town), Hubei	9/30/2008	Soil	Wang *et al.* (2010)
VX0068	23.617N	110.150E	Yulin, Guangxi	10/23/2008	Soil in banana plantation	Wang *et al.* (2010)
VX0069	30.519N	114.358E	Wuhan (CCNU), Hubei	10/24/2008	Isopod	Wang *et al.* (2010)
VX0070	30.514N	114.495E	Wuhan (Jiufeng village), Hubei	9/20/2008	Soil in cabbage field	Wang *et al.* (2010)
VX0071	23.425N	113.236E	Guangzhou (Huadu district), Guangdong	12/4/2009	Compost soil from vegetable field	Wang *et al.* (2010)
VX0072	23.425N	113.236E	Guangzhou (Huadu district), Guangdong	12/4/2009	Compost soil from vegetable field	Wang *et al.* (2010)
VX0073	24.776N	113.271E	Shaoguan (Ruyuan county), Guangdong	12/3/2009	Isopod under bamboo	Wang *et al.* (2010)
VX0074	24.776N	113.271E	Shaoguan (Ruyuan county), Guangdong	12/3/2009	Isopod under bamboo	Wang *et al.* (2010)
VX0075	23.139N	113.268E	Guangzhou (Renhe town), Guangdong	11/30/2009	Moist soil	Wang *et al.* (2010)
VX0076	25.028N	113.393E	Lechang (Changlai town), Guangdong	12/3/2009	Compost soil from vegetable field	Wang *et al.* (2010)
VX0077	25.028N	113.393E	Lechang (Changlai town), Guangdong	12/3/2009	Compost soil under bamboo	Wang *et al.* (2010)
VX0078	24.776N	113.271E	Shaoguan (Ruyuan county), Guangdong	12/3/2009	Compost soil under bamboo	Wang *et al.* (2010)
VX0079	30.519N	114.358E	Wuhan (CCNU), Hubei	8/4/2009	Isopod	Wang *et al.* (2010)

© 2012 Blackwell Publishing Ltd

neutrally evolving 10 kb sequence fragments with simcoal2 using a wide range of input mutation rates (μ, from 1×10^{-8} to 1×10^{-4} per site per generation per individual, representing $N\mu$ of 10^{-4} to 1 per site per generation) (Laval & Excoffier 2004). The FSM process follows the implementation for the 'DNA' data type in simcoal2, with 33% of mutations as transitions. Simulations incorporated a sample of 50 from a population of diploid size $N = 10\,000$, subject to no size changes or structure and a per site recombination rate of 0.000001 per site per generation per individual. We then calculated S, θ_W, π and D_{Taj} with the compute function in LDHAT2.1 (McVean *et al.* 2002), with source code edited only to eliminate manual user inputs. We then used this output to calculate $\theta_w^* = \log[L/(L - S)]/a$ (an approximate finite-sites estimator, eqn. 1 from McVean *et al.* 2002), Jukes-Cantor adjusted π, $\pi_{JC} = (-3/4) \log[1 - (4/3)\,\pi]$ (Jukes & Cantor 1969) and $D_{Taj/min} = (\pi - S/a)/|k_{min} - S/a|$, where $k_{min} = S\,(2n - 1)/n^2$ and $a = \sum_{i=1}^{n-1} \frac{1}{i}$ with sample size n and L total sites (p.166 from Schaeffer 2002) using custom perl scripts. LDHAT calculates D_{Taj} using the observed number of segregating sites.

To investigate the joint effects of population subdivision and a finite-sites mutational process, we performed similar simulations as above, except that 100 equal-sized demes were included. This analysis follows the approach of Stadler *et al.* (2009), although we use a finite sites rather than infinite-sites mutational model. We implemented a two-dimensional stepping-stone model of migration, with identical migration rates between neighbouring demes. The demes were arranged on a torus to avoid edge effects, with non-diagonal neighbours capable of exchanging migrants. We focus on this stepping-stone model of migration because it generates isolation by distance, which is found commonly in nature and which we observe in the *C.* sp. 5 data set; others have reported that island models of migration have less severe effects on patterns of polymorphism (De & Durrett 2007; Stadler *et al.* 2009). Migration rates (m) were varied from 1×10^{-6} to 1×10^{-2} per generation (i.e. Nm from 0.01 to 100), which was held constant for the most recent T generations, before that time, there was only a single population (i.e. T generations ago, a single population split instantaneously into 100 demes that then proceeded to exchange migrants until the present day). The scenarios of Stadler *et al.* (2009) considered an ancestral population of size DN splitting into D demes of size βN, with population expansion factor β. We considered an ancestral population of size N with D descendant populations of size N, essentially equivalent to the maximum expansion factor considered by Stadler *et al.* (2009) (i.e. their $\beta = D = 100$). We set T at $6.25N$, $10N$, $25N$, $100N$ and $400N$, where $N = 10\,000$ is the diploid size of each deme. We focus on this scenario for the origin of structure, which represents strong species-wide expansion owing to the colonization of new demes, as might occur following speciation or in conjunction with range expansion, in contrast to the formation of structure by fragmentation of a very large ancestral population into smaller populations (cf. Stadler *et al.* 2009). As we show, when migration rates are sufficiently high and the origin of structure is sufficiently ancient, the species-wide expansion is undetectable and the population matrix can be considered to be at equilibrium. Mutational input varied from 5×10^{-9} to 5×10^{-7} per site per generation per individual; in the limit of unconstrained migration at equilibrium, the population mutation parameter would thus be roughly $D{\cdot}N{\cdot}\mu$ (e.g. $100{\cdot}10\,000{\cdot}5 \times 10^{-7} = 0.5$). We conducted 100 replicate simulations for each parameter combination. We then computed diversity and frequency spectrum metrics as described earlier, using three different sampling schemes from the simulated sets of 100 demes: local, pooled and scattered. The local sample assessed patterns of polymorphism in sample of 50 haplotypes from a single deme; the pooled sample included 10 haplotypes sampled from each of five demes; the scattered sample included a single haplotype sampled from each of 50 demes.

To assess the effects of sample size and deme number, we performed distinct sampling on a subset of the multi-deme simulations. This subset included parameters that correspond to high and low mutation rate (5×10^{-9} and 1×10^{-7}), and to high and low migration rate (1×10^{-6} and 1×10^{-3}). First, we extracted 48 samples from the 100 demes, ranging from a fully local sample to a fully scattered sample, with eight levels of pooling (two from 24 demes, three from 16 demes, four from 12 demes, etc.). This sampling scheme holds the total sample size constant, but varies the number of demes sampled. Second, we held the number of demes sampled constant (10 demes) and varied the sample size per deme (from 1 to 10). One sample from each of 10 demes represents a scattered sample; higher sample sizes represent different extents of pooling of samples from the same number of demes. It should be noted that in all cases, simulations were conducted using the same two-dimensional stepping-stone model of migration among 100 demes described earlier.

Results

Empirical patterns of polymorphism for C. sp. 5

This sample of 10 nuclear coding loci and one mitochondrial locus identifies very high nucleotide diversity across the 41 *C.* sp. 5 strains, as well as within each

© 2012 Blackwell Publishing Ltd

Table 2 Summary of mean polymorphism and site frequency spectrum metrics for synonymous sites across 10 nuclear loci for local, pooled and scattered samples

Sample type	Sample information	*n*	Synonymous sites per locus	S_{syn} per locus	π_{syn}	π_{synJC}	D_{Taj}*	$D_{Taj/min}$
Local samples	Chengdu (Sichuan)	6	160.0	28.6	0.081	0.086	0.110	0.378
	Yaan (Sichuan)	6	156.8	24.8	0.070	0.074	0.104	0.317
	Botanical gardens (Hubei)	4	163.0	17.2	0.055	0.059	−0.425	1.868
	Jiufeng (Hubei)	4	160.4	13.0	0.046	0.048	0.488	1.822
	Lechang (Guangdong)	5	151.8	19.1	0.063	0.066	0.633	0.952
	Average across local samples	5.0	158.4	20.5	0.063	0.067	0.182	1.067
Pooled local samples	5 local samples	25	143.7	40.5	0.078	0.083	0.169	0.150
Regional pooled samples	Guangdong	11	150.8	24.0	0.055	0.057	0.100	0.139
	Hubei	11	150.3	27.0	0.063	0.067	0.077	0.151
	Sichuan	12	154.1	33.7	0.077	0.082	0.299	0.260
	Average across regional pools	11.3	151.7	28.2	0.065	0.069	0.159	0.183
Pooled all samples	All strains	41	138.8	43.4	0.075	0.080	0.145	0.125
Scattered samples	Average of 10 loci, 1000 randomized individuals	13	148.5	33.5	0.073	0.077	0.283	0.199

*Calculated using the estimated number of mutations.

sampling locality in China (Tables 1 and 2). Consistent with a previous report based on more limited sampling (Wang *et al.* 2010), we find molecular hyperdiversity in this species, with $\pi_{syn\text{-}JC}$ at nuclear loci commonly ≥5%. Across the 10 coding nuclear loci, 421 synonymous sites are bi-allelic and 17 sites segregate three different nucleotide variants. Nucleotide variation at synonymous sites in the mitochondrial COII gene is 74% higher than that of nuclear loci for all 41 strains pooled together ($\pi_{syn\text{-}JC}$ = 0.138), but only about 43% higher than nuclear loci within a given locality (mean $\pi_{syn\text{-}JC}$ across five localities = 0.0954; cf. Table 2).

To understand the degree to which molecular variability is structured across space, we partitioned our strain samples among provinces, among sampling localities and into random combinations of single samples per locality ['scattered' samples (Wakeley 1999; Lessard & Wakeley 2004)]. Polymorphism within local samples is ~19% lower than when those same samples are pooled and ~13% lower than scattered samples (Table 2). We then computed measures of structure and found moderate differentiation among provinces ($F_{st} \sim 0.14$, $G_{st} \sim 0.07$ for nuclear loci) and across localities ($F_{st} \sim 0.16$, $G_{st} \sim 0.09$) (Fig. 1; Table 3). Differentiation at the mitochondrial locus was nominally higher across localities (F_{st} = 0.27, G_{st} = 0.18) than for nuclear loci, but not among provinces (F_{st} = 0.08, G_{st} = 0.07). For the 10 nuclear loci concatenated together, we observed significant differentiation among the five localities with sample sizes ≥4 (K_s^* = 4.52, K_{st}^* = 0.045, $P < 0.0001$; F_{st} = 0.222) and among the three regions with sample sizes ≥11 (K_s^* = 4.51, K_{st}^* = 0.036, $P < 0.0001$; F_{st} = 0.187). Because high diversity can limit

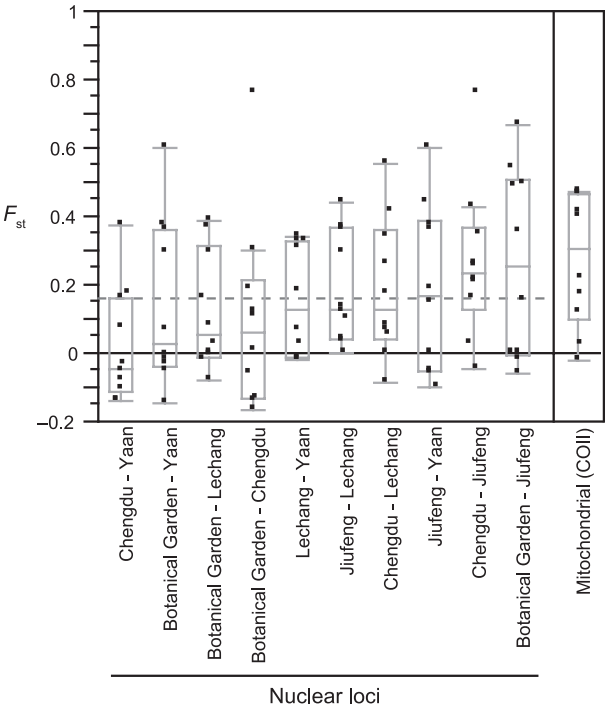

Fig. 1 Box-plots of F_{st} for pairwise contrasts of five sampling localities of *C.* sp. 5. For nuclear loci, box-plots indicate median and interquartile range among the 10 nuclear loci for each pair of local samples. Values beyond whiskers indicate potential outliers. Dashed horizontal line indicates the grand mean F_{st} across all comparisons. Box-plot for the mitochondrial locus includes all 10 pairwise comparisons of localities.

the maximum possible F_{st} and G_{st} below a value of 1 (Hedrick 2005; Jost 2008), these indices of genetic differentiation may underestimate the degree to which

© 2012 Blackwell Publishing Ltd

Table 3 Population differentiation between local and regionally pooled samples

Locus type	Sampling	Mean of means (across sample comparisons and loci)			
		F_{st}	D_a	G_{st}	G'_{st}
Nuclear	5 localities	0.162	0.00388	0.088	0.465
Nuclear	3 provinces	0.139	0.00326	0.070	0.547
Mitochondrial	5 localities	0.274	0.00928	0.176	0.824
Mitochondrial	3 provinces	0.077	0.00225	0.072	0.761

different localities have non-overlapping allele frequencies (Edelaar & Bjorklund 2011). Regardless, our aim here is simply to illustrate that there is some degree of population structure for *C*. sp. 5. This finding of moderate population structure corroborates the slight, but significant, isolation by distance reported previously in this species (Wang *et al*. 2010).

The variant frequency spectrum across loci also provides information about demographic history. A tendency for Tajima's *D* to exceed zero in a species-wide sample of *C*. sp. 5 previously was proposed to be a byproduct of population structure ('pooling') within the sample (Wang *et al*. 2010). Consequently, we quantified the skew in the frequency spectrum for our different means of partitioning strains (per-locality, pooled per-province, all strains pooled and scattered samples). Calculating D_{Taj} from the estimated number of mutations rather than the number of segregating sites partially ameliorates the skew, as seen for all samples pooled together (Fig. 2); consequently, we report D_{Taj} for *C*. sp. 5 calculated using the number of mutations in remaining figures and tables. Regardless of the way that the strains were partitioned, we consistently observed a positive skew ($D_{Taj} > 0$), indicative of a deficit of low-frequency variants relative to the expectation under a standard neutral model of molecular evolution (Table 2, Fig. 3). [The one exception is that the Wuhan Botanical Gardens sampling locality tended to have negative values of D_{Taj}.] When individual localities are considered separately, the frequency spectrum has nominally stronger skew, but not significantly more positive, than when all samples are pooled (Table 2). This pattern is perhaps consistent with the so-called 'pooling effect' that can occur more readily when populations are expanding in size (Stadler *et al*. 2009). And yet, the values of D_{Taj} are not negative on average, ruling out a strong population expansion species wide. Moreover, nuclear-locus scattered samples have highest mean D_{Taj}, so pooled samples have mean D_{Taj} that tends to be lower than both local and scattered samples (cf. Arunyawat *et al*. 2007; Stadler *et al*. 2009). The mitochondrial locus also exhibits $D_{Taj} > 0$ for all sampling schemes (local samples with mean $D_{Taj} = 1.57$, regional pool mean 0.998, scattered sample mean 0.198), but because it represents just a single locus, we focus on the sample of 10 nuclear loci. While some portion of these patterns is most likely due to the effects of population structure, the observation of a rarity of low-frequency variants even from scattered samples suggests that structure alone does not fully explain the skewed variant frequency spectrum.

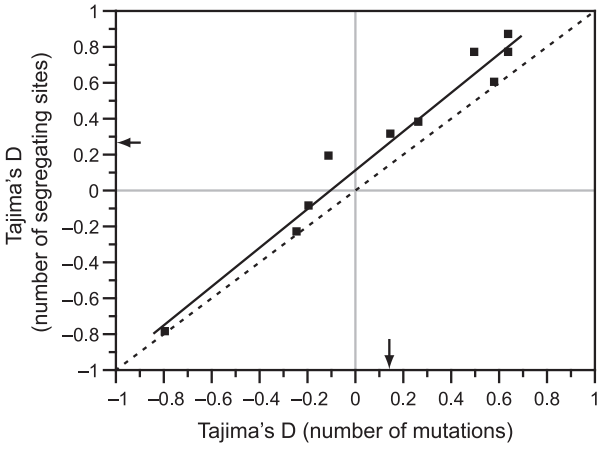

Fig. 2 Contrast of Tajima's *D* calculations for *C*. sp. 5 nuclear loci based on the number of segregating sites (*S*) vs. the estimated number of mutations. Estimates are based on the pooled sample of all 41 strains. Arrows indicate means for the 10 loci (0.267 for D_{Taj} from *S*, 0.145 for D_{Taj} from number of mutations); dotted line represents unity. Solid regression line has a significantly positive *y*-intercept (0.11, $P = 0.0136$), and a matched-pairs *t*-test indicates that D_{Taj} based on *S* is significantly greater than D_{Taj} based on the number of mutations ($t = 0.625$, d.f. = 9, $P = 0.0055$).

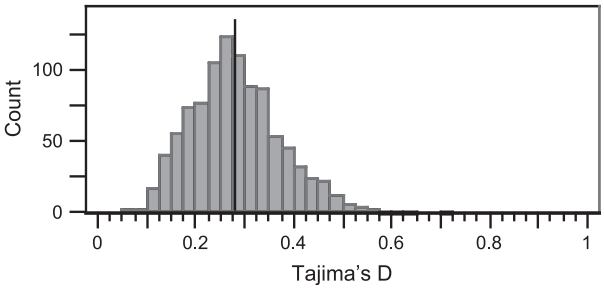

Fig. 3 Distribution of synonymous-site mean Tajima's *D* across 10 nuclear loci for scattered samples. Scattered samples are 1000 random permutations of a single strain from each of 13 sampling localities. Vertical black line indicates the mean of the distribution (0.283).

Patterns of polymorphism from coalescent simulations

Finite-sites mutation in a single panmictic population. We hypothesized that the hyperdiversity observed for

© 2012 Blackwell Publishing Ltd

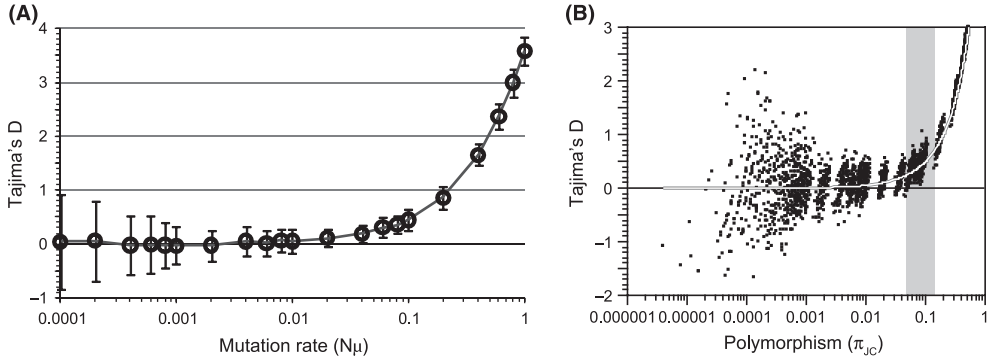

Fig. 4 Effect of increasing mutation rate (Nμ) on the site frequency spectrum (Tajima's *D*). (A) Mean Tajima's *D* as a function of input population mutation rate for coalescent simulations of a single, panmictic equilibrium population subject to a finite-sites mutation process. Error bars in (A) indicate ±1 SD. (B) Plots Tajima's *D* as a function of the computed diversity values (π_{JC}) for the simulations. White curve in (B) is a flexible spline fit. Values of $D_{Taj} > 3$ not shown in (B). The area shaded in grey in (B) indicates the biologically plausible range of π_{JC} for hyperdiverse organisms.

C. sp. 5 might skew the variant frequency spectrum, as a consequence of violation of the infinite-sites model of mutation that underlies standard expectations. We reasoned that multiple mutational hits at a given site would increase the average number of pairwise differences (π) between sequences but not increase the number of segregating sites (*S*); the difference between these two features control the sign of frequency spectrum metrics like Tajima's *D* (D_{Taj}) (Tajima 1989). Therefore, we conducted coalescent simulations under a finite-sites model of mutation for a broad range of mutation rates and computed D_{Taj}. Otherwise, the single panmictic population that we simulated conformed to standard neutral model equilibrium assumptions. We found that population mutation rates that yield $\pi_{JC} \geq 0.02$ per site do indeed inflate D_{Taj} above zero (Fig. 4). This accords with results of Bertorelle & Slatkin (1995) and of Tajima (1996). The bias towards positive values is substantial for biologically plausible mutation rates for hyperdiverse eukaryotes, with a median of $D_{Taj} \cong 0.5$ for $\pi_{JC} = 0.10$ per site. Thus, in addition to population structure and population contraction in the history of a sample being capable of causing a deficit of low-frequency variants (Tajima 1993; Stadler *et al.* 2009), high population mutation rates can induce a similar effect.

Finite-sites mutation in a subdivided population. Given that population subdivision and a finite-sites mutational process can each individually skew the variant frequency spectrum, we decided to explore these factors jointly via simulation. We know no previous treatment of the dual contributions of these processes to patterns of polymorphism, although Chikhi *et al.* (2010) considered a stepwise-mutation model of microsatellite evolution, which also involves multiple mutational hits at a locus. In our coalescent simulations, we manipulated factorial combinations of different mutation rates, migration rates and time since the origin of a structured population. We modelled the origin of population structure as an instantaneous increase of the species from a single deme of size *N* to 100 demes each of size *N*. Here, we summarize in turn the consequences of each of these parameters for estimates of polymorphism and the site frequency spectrum for local, pooled and scattered samples.

First, as expected, higher mutation rates yield higher polymorphism, regardless of sampling scheme. As for a single panmictic population, higher mutation rates also shift the site frequency spectrum towards positive values for all sampling schemes (higher D_{Taj}; Figs 5A and 6). Second, higher migration rates increase polymorphism estimates in local samples, but reduce it for pooled and scattered samples (Fig. 5B). Higher migration rates also induce an increase in D_{Taj} for local and scattered samples (Figs 5A and 6). However, more migration reduces the positively skewed site frequency spectrum in pooled samples towards zero, as also occurs for scattered samples that also have exceptionally high mutation rates (Fig. 6). Local samples are distinct from other sampling schemes in that intermediate migration rates yield more positively skewed frequency spectra than when migration is very rare or very common, as noted previously (Stadler *et al.* 2009). Local samples also exhibit much higher variance in D_{Taj} across replicate simulations than do either pooled or scattered samples.

Third, more recent origins of population structure shift D_{Taj} downward for local, pooled and scattered samples (Fig. 7). The effect of timing is most dramatic when migration is rampant, with timing having little effect on local samples when migration rates are low. For scattered samples, a recent species-wide expansion

© 2012 Blackwell Publishing Ltd

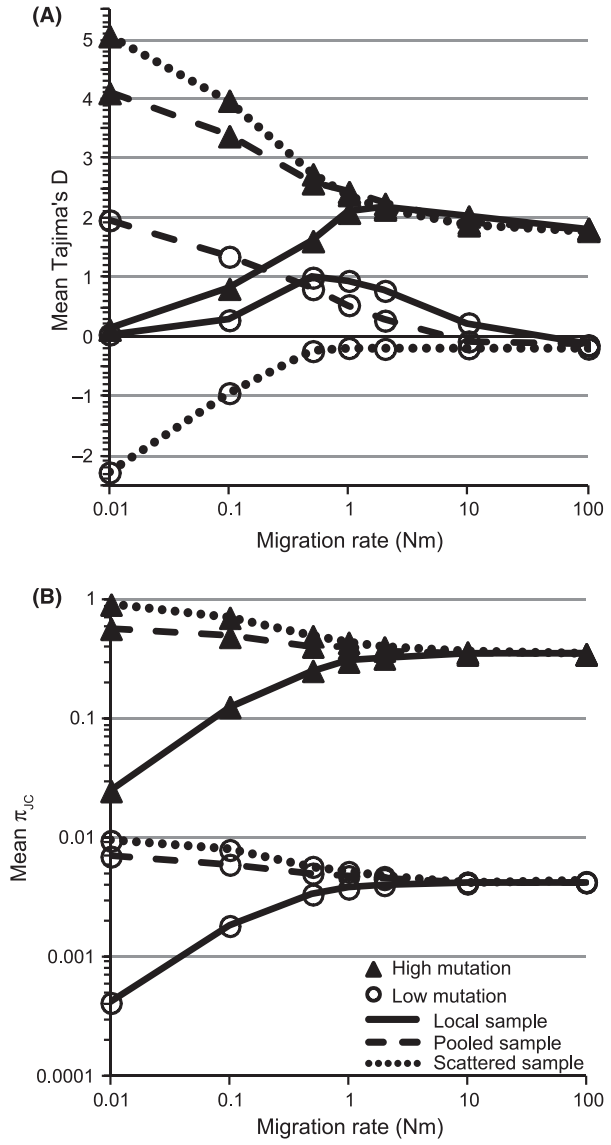

Fig. 5 Effect of sampling schemes on (A) the site frequency spectrum (Tajima's *D*) and (B) nucleotide polymorphism (π_{JC}), as a function of the migration rate (*Nm*). The legend in (B) for line dashing and icons also applies to panel (A). Finite-sites coalescent simulations modelled a 10 kb locus for 100 demes, each with population size $N = 10\,000$ connected by a 2-dimensional stepping-stone model of migration. Populations split instantaneously from a single deme $T = 100$ N generations in the past. The effect of this ancient species-wide expansion in creating an excess of low-frequency variants in scattered samples (Tajima's $D < 0$) is still apparent for low migration rates when mutation rate is low. High and low mutation rates correspond to $\mu = 5 \times 10^{-7}$ and $\mu = 5 \times 10^{-9}$, respectively.

causes negative values of D_{Taj} for all migration and mutation rates examined. Pooled samples with low migration, however, retained positive skews in the frequency spectrum, albeit shifted towards lower values than when the origin of population structure was more ancient (Fig. 7). By contrast, local samples exhibited distinct effects of a recent species-wide expansion only when migration rates were high. Indeed, with low migration among demes, the limited effect of the timing of the origin of structure yielded local samples actually producing somewhat positive values of D_{Taj} at high mutation rates (Fig. 7).

Finally, we assessed the effects of sample size and the number of demes sampled with coalescent finite-sites simulations. When we held the total sample size constant (48) and varied the number of demes sampled from one (completely local sample) to 48 (completely scattered sample), with 8 degrees of pooling between these extremes, we found that the resulting frequency spectrum summary D_{Taj} to be sensitive to the amount of pooling. Specifically, D_{Taj} (i) is more sensitive to the degree of pooling when the migration rate is low, (ii) is more strongly skewed when more samples per deme are collected, (iii) is less affected by mutation rate when fewer demes are sampled (Fig. 8A). Alternatively, we also analyzed the multi-deme populations by sampling a fixed number of 10 demes, and varied the number of samples per deme from 1 (scattered sample) to 10. This approach also shows that D_{Taj} (i) is more sensitive to the amount of pooling when the migration rate is low and (ii) is more strongly skewed when more samples per deme are analyzed (Fig. 8B).

Discussion

When a species harbours levels of nucleotide polymorphism >5% per site, such molecular hyperdiversity is more appropriately modelled with a mutation process that presumes a finite number of sites that may experience multiple mutation events in their coalescent history (Desai & Plotkin 2008; Jenkins & Song 2011). However, the ISM model forms the basis for most molecular population genetic prediction. Similarly, most conventional wisdom regarding patterns of population polymorphism arises from sampling of a single panmictic equilibrium population. Here, we investigated both the influence of FSM on the frequencies of variants in a sample, and the way that the sample of sequences was obtained from the globally structured population of a species: from a single 'local' subpopulation, from several 'pooled' subpopulations, or from a 'scattered' sample with a single representative from each of numerous subpopulations. Under a FSM process, hyperdiversity yields a null expectation substantially different from that expected for the standard neutral model with an ISM process: a seeming deficit of rare variants (positive Tajima's *D*). Differing sampling schemes generate distinct patterns of polymorphism, with higher mutation rates exacerbating deviations from the standard

© 2012 Blackwell Publishing Ltd

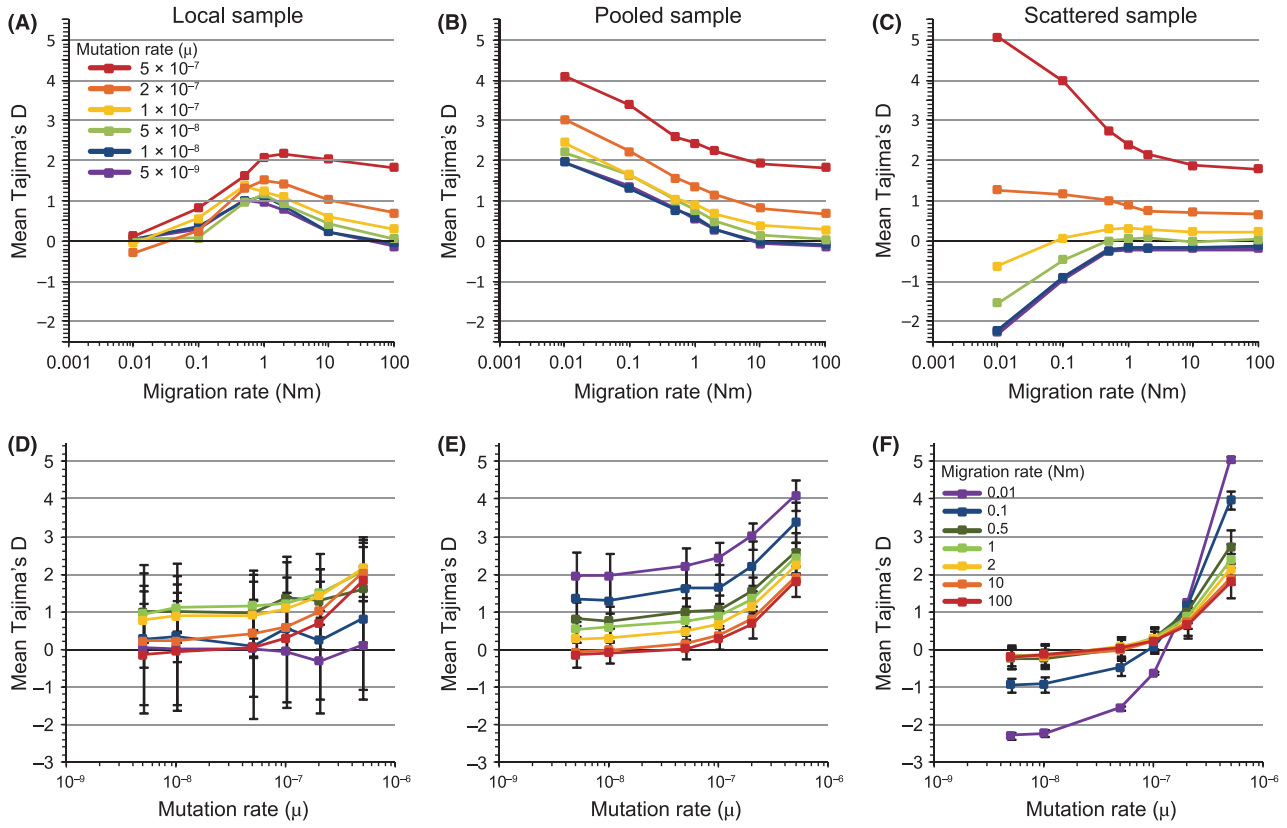

Fig. 6 The effects of migration rate (A–C) and mutation rate (D–F) on the site frequency spectrum (Tajima's *D*), calculated from local (A, D), pooled (B, E) and scattered samples (C, F). Finite-sites coalescent simulations modelled a 10 kb locus for 100 demes, each with population size N = 10 000 connected by a 2-dimensional stepping-stone model of migration. Populations split instantaneously from a single deme T = 100 N generations in the past. Error bars in the bottom panels (D–F) indicate ±1 standard deviation (variability is similar in the top panels A–C, not shown for clarity).

expectations derived previously for an ISM (Stadler *et al.* 2009). We propose that simultaneous analysis of different subsamples of a data set, representing local, pooled and scattered samples, will help clarify the demographic history of species that experience population subdivision because different sample types provide partially independent views of population history.

Finite-sites mutation processes. Following the insights of early analyses of FSMs that focused attention on the effect of mutation rate heterogeneity (Bertorelle & Slatkin 1995; Aris-Brosou & Excoffier 1996; Rogers *et al.* 1996; Yang 1996), we reasoned that multiple mutational hits at a given site would increase the average number of pairwise differences (π) between sequences but not increase the number of segregating sites (S). Because the difference between these two features controls the sign of frequency spectrum metrics like Tajima's D (D_{Taj}) (Tajima 1989), we suspected that multiple mutations in the coalescent ancestry of a population would tend to generate the appearance of a deficit of rare variants (positive D_{Taj}). Our simulations show that this is indeed the case. High population mutation rates (i.e. high polymorphism) creates a positive skew of D_{Taj} for a single panmictic population as well as for structured populations sampled according to local, pooled and scattered schemes. The skew in the site frequency spectrum is substantial when $\pi_{JC} > 5\%$, but still evident for single populations with polymorphism as low as $\pi_{JC} \cong 2\%$ per site. Thus, neutral expectations for the site frequency spectrum will experience some bias even for organisms with typical levels of molecular variation (e.g. *Drosophila* spp.). As population genomic surveys become mainstream, null models would benefit from applying FSMs to minimize this bias in large data sets. In particular, explicitly including tri-allelic nucleotide polymorphisms should help characterize the site frequency spectrum (Jenkins & Song 2011) and inference of ancestral states (De Mita *et al.* 2007).

Population structure and sampling schemes. Moreover, many species are comprised of numerous subpopulations, interconnected by migration. We considered this additional feature to ask how differing sampling

© 2012 Blackwell Publishing Ltd

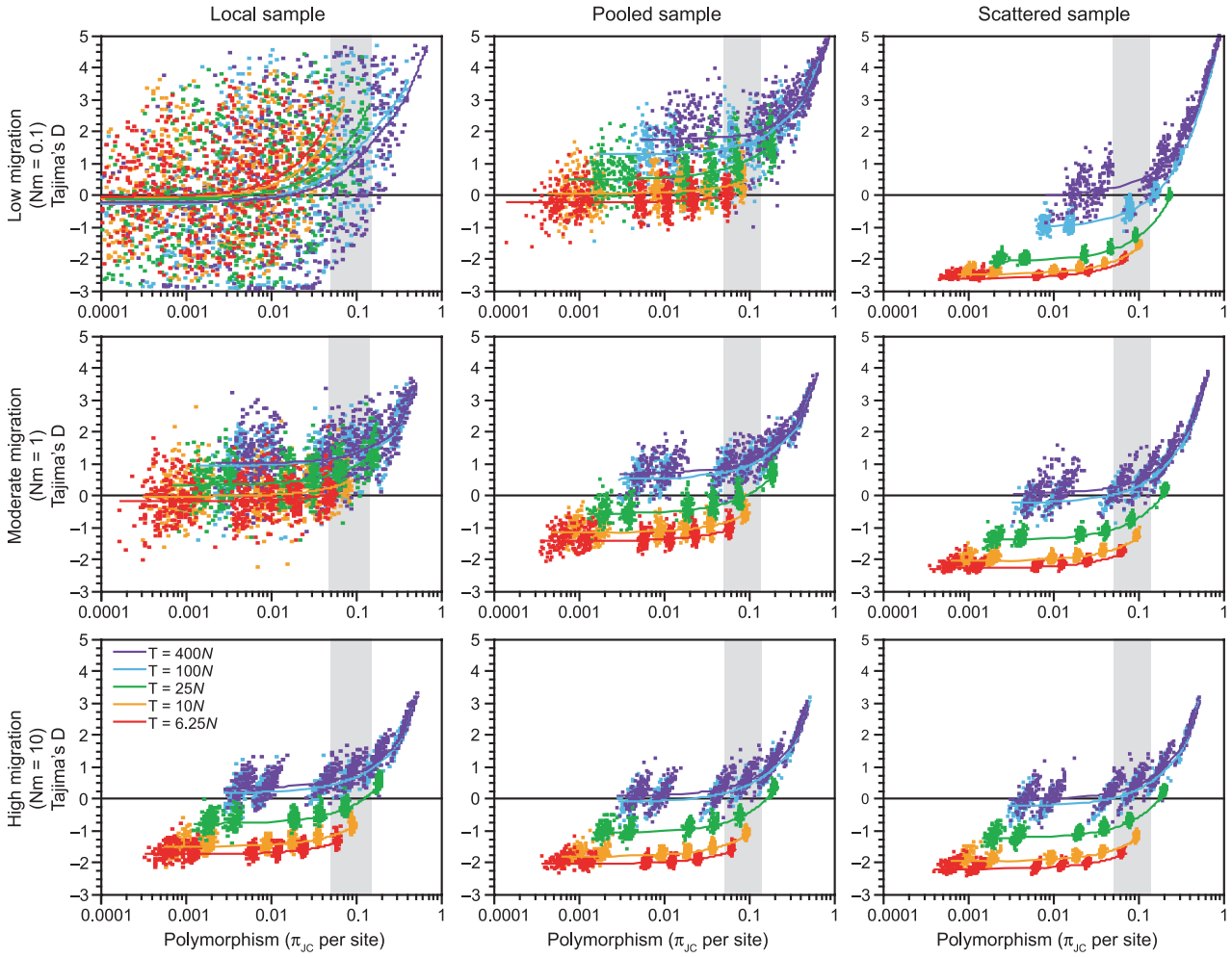

Fig. 7 The site frequency spectrum (Tajima's *D*) varies as a function of sampling scheme (local, pooled and scattered), migration rate, mutation rate and the time since the origin of population structure. Simulations model 100 demes (each with $N = 10\,000$) connected through a stepping-stone model of migration that originated from a single deme of size N instantaneously T generations in the past. Small values of T represent a recent species-wide expansion. Mutational input varies across a broad range (see Materials and Methods), yielding the π_{JC} measure of polymorphism indicated on the x-axis. Values of $\pi_{JC} < 0.0001$ are not shown. Solid curves represent flexible spline fits. The area shaded in grey indicates the biologically plausible range of π_{JC} for hyperdiverse organisms.

schemes of individuals from a species might interact with a FSM process to influence patterns of polymorphism. We simulated 100 demes connected through a stepping-stone model of migration, which generates isolation by distance, and sampled either 50 sequences from a single deme (a local sample), 10 sequences from each of five demes (a pooled sample), or one sequence from each of 50 demes (a scattered sample). These differing sampling procedures generate distinct patterns of polymorphism as a function of the population mutation rate, migration rate and the time since the origin of population structure. For example, very low migration generally induces $D_{Taj} \cong 0$ in local samples, $D_{Taj} > 0$ in pooled samples and $D_{Taj} < 0$ in scattered samples, even for relatively high rates of input of new mutations and for relatively ancient expansions of a single population into a larger structured population. By contrast, high migration rates generally induce similar values of D_{Taj} for all three sampling schemes, with ancient origins of structure inducing $D_{Taj} < 0$ and high mutational input inflating the value of D_{Taj}. The comparable patterns reported by Stadler *et al.* (2009), who considered the role of sampling strategies with an ISM model, correspond to our simulations with low mutation rates that effectively conform to an infinite-sites assumption. These results illustrate how jointly assessing patterns of polymorphism of differing sampling schemes can help dissect the demographic history of a species. Because local, pooled and scattered samples are differentially affected by mutation, migration and demographic

© 2012 Blackwell Publishing Ltd

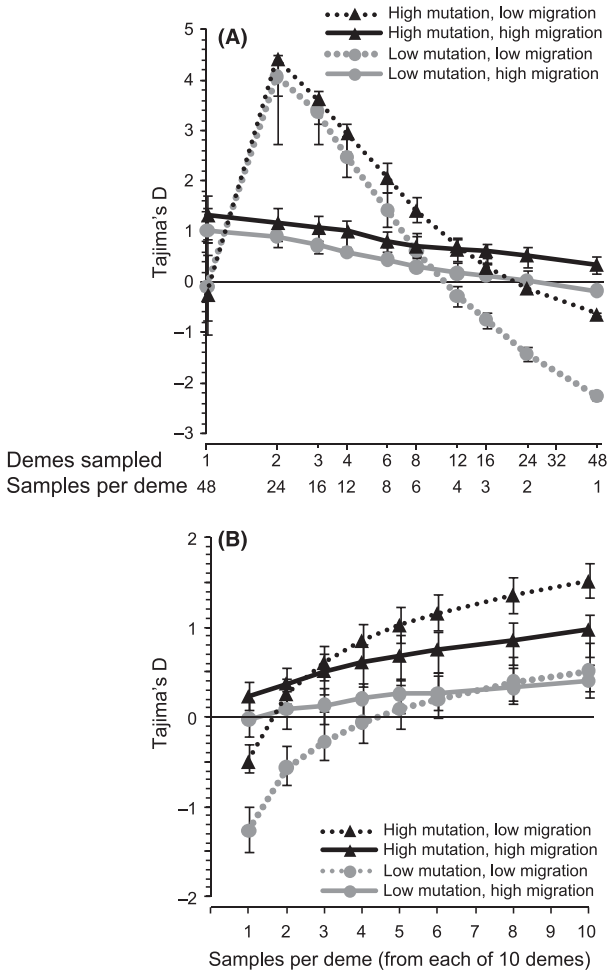

Fig. 8 Effects of the extent of pooling on the site frequency spectrum (Tajima's *D*). (A) Median Tajima's *D* (±1 quartile) when a total of 48 sequences were sampled from 1 (local sample) to 48 (scattered sample) demes. Intermediate sampling represents different degrees of pooling; note $\log_2$ scale of *x*-axis. (B) Median Tajima's *D* (±1 quartile) when samples of different sizes are taken from 10 demes (a sample of 1 corresponds to a scattered sample). High and low mutation rates correspond to $\mu = 5 \times 10^{-7}$ and $\mu = 5 \times 10^{-9}$, respectively. High and low migration rates correspond to $m = 1 \times 10^{-3}$ and $m = 1 \times 10^{-6}$, respectively. In all cases, simulations model 100 demes (each with $N = 10\,000$) connected through a stepping-stone model of migration that originated from a single deme of size *N* instantaneously $T = 100$ N generations in the past.

events, they each provide a distinct window into evolutionary history.

We used the time since the origin of structure (*T*) as a way to evaluate how different sampling strategies are affected by recent vs. ancient expansions in global population size, despite demographic stability within any given subpopulation. Our model for the origins of population structure plausibly approximate the creation of subdivision associated with range expansion or following speciation. When migration is high, all three sampling approaches respond similarly to *T*, with more recent global expansions into a structured population causing a stronger excess of low-frequency variants ($D_{Taj} < 0$). However, when migration among demes is low, the effect of *T* on the variant frequency spectrum is usually indistinguishable within local samples, and pooled samples will not typically exhibit a skew towards low-frequency variants (albeit, recent global expansions will have lower values of D_{Taj} than more ancient global expansions; Fig. 7). Moreover, the variance in D_{Taj} is very high for local samples when migration rates are low. This again illustrates how simultaneous analysis of polymorphism patterns from different sampling schemes can help demographic inference, both in terms of the extent of migration and global population size changes.

The degree of sample pooling also exerts pronounced effects on patterns of polymorphism (Fig. 8). Pooling of sequences from a small number of demes creates the strongest skew in the site frequency spectrum towards a deficit of rare variants ($D_{Taj} > 0$), with the skew becoming weaker as more demes are included in the sample. As pooling approaches the extreme of a scattered sample, with just a single representative from each deme, very low migration among the demes can induce an excess of rare variants ($D_{Taj} < 0$) under our simulation conditions. This likely reflects a signal of the origin of structure in the distant past (we simulated an instantaneous expansion of a single deme into 100 demes) that can still be captured by scattered samples when migration is sufficiently rare, whereas higher migration facilitates coalescence and effectively masks this ancient species-wide event. This change in sign of D_{Taj} as a function of the degree of pooling likely explains the distinct inferences drawn from different studies regarding the effects of sample pooling (Ptak & Przeworski 2002; Stadler *et al.* 2009).

We have concentrated primarily on how FSM and sampling schemes affect the site frequency spectrum, specifically, as summarized by Tajima's *D*. However, other population genetic characteristics also merit further consideration, such as linkage disequilibrium for different sampling schemes, the unfolded variant frequency spectrum (i.e. incorporating information on derived allele states) and the effects of FSM on estimators of differentiation like F_{st}. For example, local samples from a single deme that is embedded in a structured population tend to have greater linkage disequilibrium than would be expected from a similarly sized sample taken from single panmictic population (Wakeley & Lessard 2003; De & Durrett 2007). Appropriately characterizing the unfolded frequency spectrum with FSMs will be more challenging than when

© 2012 Blackwell Publishing Ltd

assuming an ISM, because there will be less confidence in having accurately identified ancestral and derived states when multiple mutational events at a site have occurred. A probabilistic framework of allelic states, or sticking with the folded frequency spectrum for hyper-diverse organisms, may be fruitful options. To the extent that each of these features provides independent information about the demographic history of a species, perhaps approximate Bayesian computation (ABC) methods could be extended to evaluate model fits by combining population summary statistics obtained from different sampling strategies (Beaumont *et al.* 2002; Thornton 2009). In this way, one could derive the demographic history that can best account for patterns of polymorphism across local, pooled and scattered sampling regimes, rather than restricting model fit to a particular sampling scheme. A step in this direction was implemented recently in a wild tomato data set (Tellier *et al.* 2011). Approaches that exploit distinct sampling schemes provide a complement to analyses of multidimensional variant frequency spectra based on local sampling of multiple demes for demographic modelling (Li & Stephan 2006; Gutenkunst *et al.* 2009; Siol *et al.* 2010).

Caenorhabditis sp. 5 inference. The estimates of polymorphism for *C.* sp. 5 are very similar when measured per locality, in pooled samples or in scattered samples, albeit modestly lower in local samples. Similarly, measures of the site frequency spectrum for nuclear loci are similar for both pooled and local samples, and somewhat higher on average in scattered samples (Table 2). Can a finite-sites mutational process and/or a large matrix of subpopulations connected by migration explain these patterns? We do indeed observe modest population structure and isolation by distance (Table 3) (Wang *et al.* 2010). For polymorphism to be similar and high across sampling schemes with a positive skew in the variant frequency spectrum, migration rates and mutation rates both must be high (but not the extreme mutation rates used in some of our simulations), and the population subdivision likely originated very long ago. The tendency for the mitochondrial locus to have stronger disparity in D_{Taj} across sampling schemes (local D_{Taj} > regional pool D_{Taj} > scattered D_{Taj}) might suggest that the mitochondrion experiences a lower effective migration rate than nuclear loci, potentially because of selection or male-biased dispersal.

Our simulations did not consider recent population size changes within localities. Population expansions within demes tend to depress Tajima's *D* for local samples, although scattered samples should not be much affected (Wakeley 2004); we should expect the converse for within-deme contractions. Within-deme expansions or hierarchical structure could shift D_{Taj} in different ways for different sampling schemes (K. R. St. Onge, A. E. Palme, S. I. Wright, M. Lascoux, personal communication) but appears unlikely to yield a pattern like that observed for *C.* sp. 5. Another possibility is that our attempt at creating scattered samples for *C.* sp. 5 does not recapitulate the theoretical ideal, for example, if distinct sampling localities do not properly capture independent demes. In general, appropriately designating samples as being 'scattered' presents a challenge to applying sampling scheme approaches, because different geographic localities might not correspond to distinct population genetic demes in the manner envisioned by theory (Wakeley 1999). If we have accurately assembled a scattered sample, perhaps a global population contraction could explain the tendency for D_{Taj} to be higher in scattered samples of this species, because scattered samples better reveal long-term species-wide demographic shifts. It also is conceivable that a recent increase in migration rates could cause admixture among previously more-isolated demes, which might manifest higher values of D_{Taj}. Each of these more complex types of scenarios deserves further attention.

Our analysis of *C.* sp. 5 nucleotide polymorphism has implicitly assumed that synonymous sites are selectively neutral, so that, averaged across loci, their patterns of polymorphism reflect historical population demography. However, translational selection directly affects synonymous sites in highly expressed genes of *Caenorhabditis* species, including *C.* sp. 5 (Cutter *et al.* 2008, 2009). Purifying selection on synonymous sites would tend to shift the variant frequency spectrum towards lower, more negative values, indicative of a preponderance of rare variants. To the extent that selection on synonymous sites is present, it means that the values of $D_{\mathrm{Taj}} > 0$ we observe might even be underestimates. However, only three loci included here (*Csp5-spc-1*, *Csp5-dgn-1* and *Csp5-T21B6.3*) exhibit strong codon bias, indicative of translational selection, so our inferences are unlikely to be greatly compromised by such an effect.

In some *Caenorhabditis* species, extinction-recolonization dynamics have been invoked as an important demographic process (Sivasundar & Hey 2005; Barrière & Félix 2007). However, the effects of extinction recolonization can depend on how migrants colonize new demes (Slatkin 1977). For example, with greater extinction rates, Pannell (2003) showed an increasingly negative skew of the frequency spectrum, whereas Wakeley & Aliacar (2001) reported an increasingly positive skew of the frequency spectrum. This disparity likely reflects a sensitivity to assumptions about the process of colonization of new demes, either via a 'propagule pool' of *k*

© 2012 Blackwell Publishing Ltd

individuals from a single parent deme (Pannell 2003) or via a 'migrant pool' of k individuals each from different parent demes (Wakeley & Aliacar 2001). Colonization by a few individuals from a single parent deme likely will induce a founder bottleneck, leading to the typical 'population expansion' signature of an excess of rare variants ($D_{\mathrm{Taj}} < 0$ for local samples). By contrast, founding of new demes by migrants from a small number (>1) of parent demes should generate patterns of polymorphism similar to admixture or to pooled samples ($D_{\mathrm{Taj}} > 0$ for local samples). If extinction recolonization is an important feature of *C.* sp. 5 demography, then the general deficit of low-frequency variants that we observe in local samples is more consistent with colonization akin to a migrant pool dispersal mechanism.

Overall, we demonstrate that patterns of polymorphism for *C.* sp. 5 are largely concordant across distinct sampling schemes. This result is consistent with high migration among localities, coupled with a high population mutation parameter (likely driven by a large effective population size). Moreover, the patterns across sampling schemes are inconsistent with a recent species-wide population expansion involving the creation of many new subpopulations or with extinction-recolonization dynamics involving propagule-pool dispersal. *C.* sp. 5 provides an example of the potential to more completely evaluate population history by jointly assessing population genetic variation from local, pooled and scattered samples.

Acknowledgements

We are grateful for the assistance of Richard Jovelin and Alivia Dey, as well as constructive discussions with Stephen Wright. GXW is supported by the National Natural Science Foundation of China and HuaZhong Normal University (#31071998, 30870350). ADC is supported by the NIH, a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada and by a Canada Research Chair.

References

Aris-Brosou S, Excoffier L (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution*, **13**, 494–504.

Arunyawat U, Stephan W, Stadler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution*, **24**, 2310–2322.

Bachtrog D, Andolfatto P (2006) Selection, recombination and demographic history in *Drosophila miranda*. *Genetics*, **174**, 2045–2059.

Barrière A, Félix M-A (2006) Isolation of *C. elegans* and related nematodes. In: *WormBook* (ed. The *C. elegans* Research Community) . WormBook, DOI:10.1895/wormbook.1.115.1.

Barrière A, Félix M-A (2007) Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. *Genetics*, **176**, 999–1011.

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

Bertorelle G, Slatkin M (1995) The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. *Molecular Biology and Evolution*, **12**, 887–892.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, **186**, 983–995.

Cutter AD (2008) Multilocus patterns of polymorphism and selection across the X-chromosome of *Caenorhabditis remanei*. *Genetics*, **178**, 1661–1672.

Cutter AD, Wasmuth JD, Washington NL (2008) Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. *Genetics*, **178**, 2093–2104.

Cutter AD, Dey A, Murray RL (2009) Evolution of the *Caenorhabditis elegans* genome. *Molecular Biology and Evolution*, **26**, 1199–1234.

De A, Durrett R (2007) Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics*, **176**, 969–981.

De Mita S, Ronfort J, McKhann HI *et al.* (2007) Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in *Medicago truncatula*. *Genetics*, **177**, 2123–2133.

Desai MM, Plotkin JB (2008) The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, **180**, 2175–2191.

Edelaar P, Bjorklund M (2011) If Fst does not measure neutral genetic differentiation, then comparing it with Qst is misleading. Or is it?. *Molecular Ecology*, **20**, 1805–1812.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*, **5**, e1000695.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Hudson RR (2000) A new statistic for detecting genetic differentiation. *Genetics*, **155**, 2011–2014.

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, **9**, 138–151.

Jenkins PA, Song YS (2011) The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology*, **80**, 158–173.

Jost L (2008) Gst and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism* (ed. Munro HN), pp. 21–132. Academic Press, New York.

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.

© 2012 Blackwell Publishing Ltd

Lessard S, Wakeley J (2004) The two-locus ancestral graph in a subdivided population: convergence as the number of demes grows in the island model. *Journal of Mathematical Biology*, **48**, 275–292.

Li HP, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genetics*, **2**, 1580–1589.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

Nei M, Li WH (1979) Mathematical-model for studying genetic-variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences United States of America*, **76**, 5269–5273.

Pannell JR (2003) Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution*, **57**, 949–961.

Perez-Losada M, Browne EB, Madsen A *et al.* (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infection Genetics and Evolution*, **6**, 97–112.

Ptak SE, Przeworski M (2002) Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, **18**, 559–563.

Ray N, Currat M, Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution*, **20**, 76–86.

Rogers AR, Fraley AE, Bamshad MJ, Watkins WS, Jorde LB (1996) Mitochondrial mismatch analysis is insensitive to the mutational process. *Molecular Biology and Evolution*, **13**, 895–902.

Schaeffer SW (2002) Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura*. *Genetics Research*, **80**, 163–175.

Siol M, Wright SI, Barrett SCH (2010) The population genomics of plant adaptation. *New Phytologist*, **188**, 313–332.

Sivasundar A, Hey J (2005) Sampling from natural populations using RNAi reveals high outcrossing and population structure in *Caenorhabditis elegans*. *Current Biology*, **15**, 1598–1602.

Slatkin M (1977) Gene flow and genetic drift in a species subject to frequent local extinctions. *Theoretical Population Biology*, **12**, 253–262.

Stadler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, **182**, 205–216.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tajima F (1993) Statistical analysis of DNA polymorphism. *Japanese Journal of Genetics*, **68**, 567–595.

Tajima F (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, **143**, 1457–1465.

Tellier A, Laurent SJ, Lainer H, Pavlidis P, Stephan W (2011) Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 17052–17057.

Thomas WK, Wilson AC (1991) Mode and tempo of molecular evolution in the nematode *Caenorhabditis*: cytochrome oxidase-II and calmodulin sequences. *Genetics*, **128**, 269–279.

Thornton K (2009) Automating approximate Bayesian computation by local linear regression. *BMC Genetics*, **10**, 35.

Wakeley J (1999) Nonequilibrium migration in human history. *Genetics*, **153**, 1863–1871.

Wakeley J (2004) Metapopulation models for historical inference. *Molecular Ecology*, **13**, 865–875.

Wakeley J, Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893–905.

Wakeley J, Lessard S (2003) Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics*, **164**, 1043–1053.

Wang G-X, Ren S, Ren Y, Ai H, Cutter AD (2010) Extremely high molecular diversity within the East Asian nematode *Caenorhabditis* sp. 5. *Molecular Ecology*, **19**, 5022–5029.

Yang Z (1996) Statistical properties of a DNA sample under the finite-sites model. *Genetics*, **144**, 1941–1950.

Research in A.C.'s laboratory addresses genome evolution, reproductive isolation, and adaptation, focusing on population genetic and molecular evolutionary approaches in nematodes. G.-x.W.'s lab studies population genetics and molecular evolution in Caenorhabditis and entomopathogenic Mermithid nematodes. Y.P. and H.A.'s research interests in the Wang lab include studying the biodiversity and phylogenetics of nematodes in China.

Data accessibility

All new sequences are available in Genbank under accession numbers JQ219362–JQ219658, including the strain identifiers indicated in Table 1. Sequences from Wang *et al.* (2010) for these strains are available in Genbank under accession numbers HQ189144–HQ189296.

© 2012 Blackwell Publishing Ltd