Genet. Res., Camb. (2013), **95**, *pp.* 57–67.
doi:10.1017/S0016672313000062

Stochastic patterns of polymorphism after a selective sweep over a subdivided population

YUSEOB KIM*
Division of EcoScience, Ewha Womans University, Seoul 120-750, Korea

(*Received 6 November 2012; revised 8 March 2013; first published online 8 April 2013*)

Summary

The geographic structure of a population, which is modelled as a network of several small random-mating populations or demes exchanging migrants between them, limits the rapid spread of a beneficial allele under strong directional selection to the entire population. This weakens or modifies the hitchhiking effect of the beneficial allele on the pattern of genetic variation at linked neutral loci. Previous studies suggested that the characteristic patterns of polymorphism arise with selective sweeps in such a subdivided population. However, they did not fully address the stochastic pattern, as expected in an actual sample of DNA sequence, of such patterns. This study uses a novel method of individual-based forward-in-time simulation to generate multi-locus neutral polymorphism after a selective sweep in a moderately subdivided population. Population subdivision is shown to cause frequency spectrum to shift slightly such that Tajima's D becomes less negative than expected under a panmictic population. Similarly, the pattern of linkage disequilibrium showed very small change due to population subdivision. On the other hand, the value of Wright's F_{ST} at closely linked neutral loci relative to that at unlinked loci greatly increased by population subdivision as predicted by previous studies. Finally, the distribution of the gradient of heterozygosity along the migration path of beneficial mutation, previously suggested to allow the inference of the direction of spread, was investigated. The variance of difference in heterozygosity was much larger than the mean, suggesting that such an inference may not be practical.

1. Introduction

Characteristic changes in DNA sequence polymorphism, such as a locus-specific reduction in the amount of variation (Maynard Smith & Haigh, 1974; Kaplan *et al*., 1989), skew in site-frequency spectrum (Braverman *et al*. 1995; Fay & Wu, 2000), and characteristic patterning of linkage disequilibrium (LD) (Kim & Nielsen, 2004; Stephan *et al*., 2006), are caused by the fixation of an advantageous allele driven by strong directional selection. This effect, termed selective sweep or genetic hitchhiking, provides a strong evidence of recent adaptive evolution at a local genomic region that is otherwise difficult to prove. Advances in the mathematical theory of selective sweeps allowed researchers to identify the characteristic signature from DNA sequence polymorphism, statistically distinguishable from randomly generated patterns under genetic drift, and obtain basic information about adaptation such as the intensity of selection. However, detailed predictions on the pattern of DNA sequence polymorphism provided by the current theory of selective sweeps may not be robust to the standard assumptions of the simple demographic structure of population (e.g. large random mating population) and the mode of selection (e.g. constant selective pressure on a newly arising mutation). At the same time, a deviation of the sweep pattern from the standard prediction may contain information about biological details in the complex nature of adaptive evolution.

Individuals in a natural population are distributed over a geographical space and this structure of population is often modelled as a network of subpopulations or demes, within which individuals mate randomly. The pattern of selective sweeps in spatially subdivided populations was investigated in several studies (Slatkin & Wiehe, 1998; Wiehe *et al*., 2005; Faure *et al*., 2008; Bierne, 2010; Kim & Maruki, 2011). They focused on how the measure of population differentiation, Wright's F_{ST} statistic, is affected by a selective sweep spreading across demes.

* Corresponding author: Division of EcoScience, Ewha Womans University, Seoul 120-750, Korea. E-mail: yuseob@ewha.ac.kr

Their results suggest that, if populations are already highly differentiated, a sweep of a common haplotype linked to the selected allele leads to a local reduction in genetic differentiation (decrease in F_{ST}) (Santiago & Caballero, 2005). On the other hand, if populations are initially weakly differentiated (low F_{ST}), the stochastic breakdown of association between the selected allele and neutral alleles by recombination leads to a local increase in F_{ST} (Slatkin & Wiehe, 1998; Bierne, 2010). Investigation on this effect was further extended in Kim & Maruki (2011), which mainly considered a subdivided population with very frequent migration among demes ($m << 1$ but $Nm > 1$, where m is the migration rate per lineage per generation and N is the effective population size of a deme) and thus the spatial population structure is not 'visible' by F_{ST} measured at random neutral loci. Still in this case, the spread of beneficial allele, with selective advantage s, over the entire population is slower than in a completely panmictic population as long as $m < s$. This delay in the propagation of the beneficial allele provides more opportunities for the breakdown of association between the beneficial and neutral alleles, thus leading to weaker signature of selective sweep. At the same time, the patterns of polymorphism after sweep become spatially structured: around the selective target locus, F_{ST} increases as predicted by Slatkin & Wiehe (1998). Furthermore, the expected amount of variation (heterozygosity) is lowest in the first deme (where the beneficial mutation first occurred and started) but gradually increases as distance from the first deme, along the path of beneficial mutants' spread, increases (Kim & Maruki, 2011).

When the spatial structure of population causes deviations in the pattern of selective sweep, compared with prediction by the standard model of selective sweep, as described above, it may affect the statistical inference and parameter estimation of positive selection from DNA sequence data if the statistical method is based on a standard model assuming panmixia. On the other hand, such deviations may provide information on biological details in the process of adaptive evolution. For example, by observing the gradient of heterozygosity left after sweep, one may infer the spatial origin of beneficial mutation (Kim & Maruki, 2011). However, it is not known whether such a pattern can be confidently inferred in actual genetic data, which is expected to be greatly influenced by stochastic noises. For these reasons, it is important to predict how clearly the unique patterns of selective sweep would be identified in DNA sequences sampled over a subdivided population. However, previous studies listed above mainly focused on the expectation, not the stochastic distribution, of changes in heterozygosity and F_{ST}. Furthermore, previous results are for one neutral locus near the location of selection. The patterns of polymorphism at multiple linked neutral loci – such as frequency spectrum and LD summarized by statistics like Tajima's D and r^2 (Hill & Robertson, 1968; Tajima, 1989) – are likely to contain more information than single-site heterozygosity and F_{ST}. This study therefore aims to predict full stochastic patterns of multi-locus polymorphism in DNA sequences sampled over a subdivided population. Here, as in Kim & Maruki (2011), hard and complete selective sweeps are considered: the beneficial mutation occurs once and increases due to selection until it is fixed. A different model of directional selection on continuous geographic space, resulting in the pattern of soft selective sweep due to the spread of multiple beneficial alleles, was investigated in Ralph & Coop (2010).

As the multi-locus stochastic model of selective sweep is mathematically intractable, this study examines the pattern using computer simulation. Coalescent simulations are commonly used to obtain the distribution of Tajima's D and other sample statistics. Although it is very fast to allow the exploration of multi-dimensional parameter space, coalescent simulation is however feasible for only limited sets of demographic and selective scenarios. For the complicated (but realistic) models of demography and selection, individual-based forward-in-time simulation is more straightforward and feasible without requiring advanced programming skills. The notorious problem of low speed in individual-based simulations can be solved for the simulation of selective sweeps, as outlined in Kim & Wiehe (2009). Namely, evolutionary process in the whole population is simulated forward-in-time only for the selective phase, which is defined as the period from the birth to the fixation of the beneficial mutation. Then, the structure of genealogy (coalescent tree) during the selective phase is extracted and then combined to standard neutral polymorphism expected at the start of selective phase. (Here, combining pre-determined ancestral polymorphism with individual-based forward simulation should not be confused with another simulation method that constructs the trajectory of selected allele forward in time and then builds genealogy at linked loci using the principle of structured coalescent (Spencer & Coop, 2004; Ewing & Hermisson, 2010).) This approach led to very fast estimation of expected heterozygosity in complex models of selection (e.g. Kim & Stephan, 2003) without explicitly modelling mutation processes. For the current investigation, this simulation method is further extended to generate the multi-locus pattern of bi-allelic polymorphism as expected for actual DNA sequence data. The method developed here can be applied to any other complex models of selective sweeps.

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

2. Model and simulation methods

This study considers a population with N haploid individuals that is subdivided into ten demes of equal sizes ($N_d = N/10$). Following Kim & Maruki (2011), these demes are spatially arranged according to a circular stepping-stone model. Demes are indexed by 1–10 according to their spatial arrangement. Reproduction occurs in discrete generations according to the Wright–Fisher model to which the steps of selection, recombination and migration are added. Recombination is assumed to occur by a random union of two haploids followed by meiosis. The evolutionary dynamics studied here is therefore equivalent to that of a population with $N/2$ diploid individuals under additive selection (no dominance). During migration, a given haploid in deme k moves to deme $k-1$ (10 if $k=1$) or $k+1$ (1 if $k=10$) with probability m. A mutation to a beneficial allele, denoted B, with selection coefficient s occurs in deme 1 and this allele spreads to the entire population by positive selection.

To investigate the stochastic patterns of selective sweeps at multiple neutral loci, this study performs an individual-based simulation in which a haploid individual is represented by a chromosome with 200 evenly spaced neutral loci and the locus of the beneficial allele (denoted 'B locus') that is located between the 100th and 101st neutral loci. A neutral locus here corresponds to a short DNA segment that harbours at most one polymorphic site as observed in a set of sampled DNA sequences (see below). Recombination occurs between adjacent neutral loci with probability r_n (with $r_n/2$ between the B locus and an adjacent neutral locus). While individual-based forward-in-time simulations that keep track of bi-allelic evolutionary dynamics at multiple neutral loci are generally too slow to be practical in exploring multi-dimensional parameter space, this study uses an approach to shorten the simulation time by extracting the structure of genealogy at all loci during the selective phase (time between the occurrence and the fixation of the advantageous allele) as described in Kim & Wiehe (2009).

Simulation starts with one chromosome carrying the B allele at time $t=0$ in deme 1 and other $N-1$ chromosomes carrying the ancestral allele b. To trace gene lineages at each neutral locus forward in time, chromosomes in the population at $t=0$ are indexed by distinct 'ancestral numbers' (1, …,N). Namely, all of the N neutral lineages at a locus are distinctly marked. In the subsequent generations (forward-in-time) many of these lineages are lost by genetic drift or by hitchhiking effect. The B allele may also be lost by genetic drift. In that case, simulation starts again from the initial condition described above. Let τ be the number of generations it takes for the B allele to be fixed in the entire population. Considering one neutral locus, let $p_i(t)$ be the frequency of ancestral number i ($=1, \ldots, N$) at time t during a simulation run. Hence, $p_i(0) = 1/N$ and $p_1(t) + p_2(t) + \cdots + p_N(t) = 1$ for all t. With strong selection, the duration of a selective sweep is very short in the time scale of neutral mutation and coalescent. Therefore, new neutral mutations between time 0 and τ can be ignored. Then, the expected heterozygosity after selective sweep relative to that before sweep is given by $1-\{p_1(t)^2 + p_2(t)^2 + \cdots + p_N(t)^2\}$ (Kim & Wiehe, 2009). Note that mutation process is not explicitly modelled to obtain this result. However, in order to investigate the multi-site stochastic patterns of variation as expected in actual DNA sequence polymorphism, it is necessary to generate the bi-allelic polymorphism by modelling explicit mutational products (alleles defined by identity by state) in the simulation. This is done as follows.

First, at time $t=\tau$, n chromosomes are sampled over demes according to the sampling scheme described below. Then, lineages at each neutral locus are traced by identifying ancestral numbers that chromosomes carry. Let n_0 be the count of distinct ancestral numbers at the locus observed in the sample. This count represents how many distinct lineages ancestral to the sample are present at time $t=0$. For example, if strong hitchhiking effect causes all neutral lineages to coalesce (looking backward in time) during the selective phase, $n_0 = 1$. If $n_0 > 1$ and it is assumed that the population evolved at neutral equilibrium before the start of selective sweep, the expected genealogy before sweep is given by the standard neutral coalescent tree starting with n_0 lineages. Therefore, n_0 determines the overall size of the genealogy (thus the expected level of variation in the sample) at the locus. Since new mutations between $t=0$ and τ are ignored, a locus in the sample is polymorphic only if $n_0 > 1$ and if there was already polymorphism among the n_0 sequences at time 0. The probability of the latter is given by standard neutral theory: a derived allele is carried on k randomly chosen lineages with probability $\theta/k = 2N\mu/k$ ($k = 1, \ldots, n_0-1$; $\theta << 1$) where μ is the mutation rate per neutral locus per generation. Then, the descendants of these k lineages in the present sample receive the derived allele. Define $a(n) = 1 + 1/2 + 1/3 + \cdots + 1/(n-1)$. The n_0 ancestral sequences are monomorphic at the locus ($k=0$) with probability $1-\theta a(n_0)$. In the simulation, k is drawn for each locus from the above distribution with $\theta = 1/a(n)$. Therefore, if $n_0 = n$ (no coalescence among lineages during the selective phase), the sample always harbour polymorphism at the locus. If $n_0 < n$, polymorphism is generated with probability $a(n_0)/a(n)$. In this way, mutations are placed over loci proportional to their expected sizes of genealogy.

Replicates of simulated data are generated in two steps. First, this individual-based simulation

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

from time 0 to τ is performed to generate one realization of a population at the end of sweep. For this result, the procedure of random sampling n chromosomes and then assigning the ancestral polymorphism at $t=0$ is repeated K_S times to generate K_S samples of multi-locus, bi-allelic polymorphism. This procedure is repeated K_W times (generating K_W whole-population replicates). In total, $K_W \times K_S$ replicates of samples are generated. In most cases, $K_W=200$ and $K_S=5$ were used. Results obtained with $K_W=1000$ and $K_S=1$ were not qualitatively distinguishable (Table 1), suggesting that potential correlation among K_S replicates extracted from one realization of whole-population genealogy is not a problem.

Note that the above procedure of modelling polymorphism in n_0 ancestral lineages at a locus constrains the scope of scenarios to be simulated: before the start of selective sweep ($t=0$), the ancestral population is not subdivided (genetically differentiated). It is because, regardless of which demes the n chromosomes are sampled from, the n_0 ancestral lineages are assumed to enter the process of neutral coalescent in a single panmictic population of size N. This assumption is justified if the scaled migration rate $N_d m$ is sufficiently large, so that the neutral coalescent in the stepping-stone model is effectively identical to that in a panmictic population. Or it simulates the scenario in which the ancestral panmictic population is split into ten demes immediately before selective sweep starts. To partially overcome this restriction on scenarios, population differentiation at the start of selective sweep was artificially introduced by running the forward simulation without any migration between demes for L_I generations before beneficial mutation occurs in deme 1 (see below for more details).

The simulation procedure above also effectively imposes a uniform density of polymorphic sites on chromosomes that are ancestral to the sample. In reality, the ancestral polymorphism would not be uniformly distributed because mutation process is Poisson and genealogies at adjacent loci are correlated due to partial linkage. Therefore, the current simulation generates less heterogeneity in polymorphic site density than would be observed in an actual sample. However, the focus of this study is to examine the stochastic pattern of frequency spectrum, LD, and the spatial (geographic) patterning of heterozygosity. The distributions of these quantities are not expected to be sensitive to the heterogeneity of polymorphic site density. Uniform ancestral polymorphism across sites may also have the effect of so called 'fixed S scheme' (where S is the pre-determined number of mutations mapped on genealogy in a coalescent simulation) that moderately distorts the distribution of summary statistics as pointed out by Wall & Hudson (2001). This problem however originates due to failure in mapping mutation events proportional to the length of genealogy: while the size of coalescent tree is highly variable, a fixed number of mutations are placed. The current method is not likely to suffer such a problem, as the pattern of ancestral polymorphism is not assigned given any particular realization of genealogy at the site. In addition, as n_0 is much smaller than n in most sites due to hitchhiking effect, the placements of ancestral polymorphism are made proportional to the (variable) sizes of genealogy across sites, greatly reducing the potential effect of fixed S scheme.

3. Results

(i) *Parameter values*

Simulation was performed to generate a sample of n chromosomes when the beneficial mutation has completed its spread over ten demes in the circular stepping-stone model. The exploration of parameter space begins with $N=50\,000$ ($N_d=5000$ for each deme), $s=0{\cdot}1$ and $r_n=10^{-4}$. As there are 100 neutral loci on either side of the B locus (site under selection), scaled recombination rate r/s between the most distal locus and the B locus is approximately 0·1. It is known that important signatures of hitchhiking (e.g. negative Tajima's D, peak of LD, change in F_{ST} in a subdivided population) are contained within such a range of r/s (Kaplan *et al.*, 1989; Braverman *et al.*, 1995; Slatkin & Wiehe, 1998; Stephan *et al.*, 2006; Bierne, 2010; Kim & Maruki, 2011). With this basic parameter set, the effect of population subdivision was examined by varying migration rate m (m/s ranging from 0·01 to 3) and the scheme of sampling chromosomes over different demes. Let n_k be the number of chromosomes sampled from deme k. The sampling scheme is denoted by $\{(i, n_i), (j, n_j), \ldots, (k, n_k)\}$, where $i, j, \ldots, k$ are demes from which at least one chromosome is sampled. For each parameter set, simulation generated at least 1000 samples. Then, the distributions of various sample statistics were obtained to examine the effect of population subdivision on the frequency spectrum, LD and the gradient of heterozygosity.

(ii) *Frequency spectrum*

The site frequency spectra for the samples of 20 chromosomes were observed at neutral loci with various distances to the B locus with varying migration rate (Fig. 1). The overall modification of frequency spectrum due to population subdivision (decreasing m) appears to be minor. However, the proportion of sites with intermediate-frequency derived alleles is shown to increase with decreasing m.

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

Table 1. *Results of simulating selective sweeps in subdivided population* (*mean* ± SD)

Case	r_n[a]	m	L_I[b] (F_{ST})[c]	D[d]	$r^{2(L)}$	ω	F_{ST}[e]	$> Q_{0\cdot99}$[f]	$\Delta\pi$	$> Q_{0\cdot99}$[g]
1	10^{-4}	0·3	0	−0·69 ± 0·38	0·116 ± 0·022	4·52 ± 1·35	0·0032 ± 0·0093	0·028	0·0062 ± 0·097	—
2	10^{-4}	0·01	0	−0·49 ± 0·40	0·115 ± 0·021	4·09 ± 1·13	0·030 ± 0·022	0·225	0·10 ± 0·12	0·096
3[h]	10^{-4}	0·01	0	−0·46 ± 0·41	0·115 ± 0·021	4·05 ± 1·08	0·032 ± 0·025	0·262	0·10 ± 0·12	0·095
4	10^{-4}	0·01	100 (0·025)	−0·41 ± 0·37	0·115 ± 0·021	4·01 ± 1·03	0·033 ± 0·025	0·284	0·11 ± 0·12	0·102
5	10^{-4}	0·01	400 (0·059)	−0·43 ± 0·41	0·116 ± 0·021	4·17 ± 1·17	0·034 ± 0·027	0·281	0·098 ± 0·12	0·094
6	10^{-4}	0·001	0	0·010 ± 0·43	0·126 ± 0·027	3·68 ± 1·17	0·127 ± 0·071	0·733	0·096 ± 0·15	0·145
7[h]	10^{-4}	0·001	0	−0·021 ± 0·45	0·126 ± 0·028	3·63 ± 1·02	0·129 ± 0·072	0·744	0·096 ± 0·17	0·175
8	10^{-4}	0·001	100 (0·025)	0·060 ± 0·44	0·125 ± 0·026	3·53 ± 1·00	0·137 ± 0·075	0·790	0·074 ± 0·15	0·113
9	10^{-4}	0·001	400 (0·060)	0·041 ± 0·46	0·130 ± 0·027	3·63 ± 0·98	0·125 ± 0·065	0·764	0·088 ± 0·16	0·159
10	5×10^{-5}	0·3	0	−1·30 ± 0·45	0·211 ± 0·071	10·9 ± 12·7	0·0029 ± 0·012	0·061	0·010 ± 0·17	—
11	5×10^{-5}	0·01	0	−1·02 ± 0·55	0·207 ± 0·068	9·38 ± 11·9	0·030 ± 0·032	0·248	0·12 ± 0·20	0·061
12	5×10^{-5}	0·001	0	−0·48 ± 0·72	0·230 ± 0·091	7·65 ± 8·39	0·139 ± 0·109	0·636	0·11 ± 0·28	0·125
13[i]	$2{\cdot}5 \times 10^{-5}$	0·3	0	−1·73 ± 0·62	0·399 ± 0·192	74·6 ± 129	0·0011 ± 0·013	0·062	0·0059 ± 0·29	—
14[i]	$2{\cdot}5 \times 10^{-5}$	0·01	0	−1·53 ± 0·68	0·387 ± 0·185	56·9 ± 98·9	0·027 ± 0·040	0·210	0·17 ± 0·31	0·039
15[i]	$2{\cdot}5 \times 10^{-5}$	0·001	0	−0·94 ± 0·93	0·427 ± 0·208	42·2 ± 115	0·129 ± 0·127	0·544	0·12 ± 0·40	0·065

Other parameters: $N = 50\,000$, $s = 0{\cdot}1$, sample = {(2, 20), (5, 20)}, $K_W = 200$, $K_S = 5$.
[a]Per-locus recombination rate.
[b]The initial length of simulation without any migration between demes before the start of selective sweep.
[c]Mean F_{ST} between demes 2 and 5 at the start of selective sweep.
[d]Tajima's D.
[e]Mean F_{ST} between demes 2 and 5 after the completion of selective sweep.
[f]Proportion of replicates where F_{ST} between demes 2 and 5 is greater than the 99th percentile from neutral simulations corresponding to cases 1, 2 and 6 (0·0250, 0·0420, 0·0756 for $m = 0{\cdot}3$, 0·01, 0·001, respectively).
[g]Proportion of replicates $\Delta\pi$ is greater than the 99th percentile ($Q_{0\cdot99}$) from panmictic population. For cases 2–9, $Q_{0\cdot99} = 0{\cdot}253$ (from case 1); for cases 11 and 12, $Q_{0\cdot99} = 0{\cdot}433$ (from case 10); for cases 14 and 15, $Q_{0\cdot99} = 0{\cdot}724$ (from case 13).
[h]Using $K_W = 1000$ and $K_S = 1$.
[i]Replicates with at least three polymorphic sites at each half of the chromosome were considered.

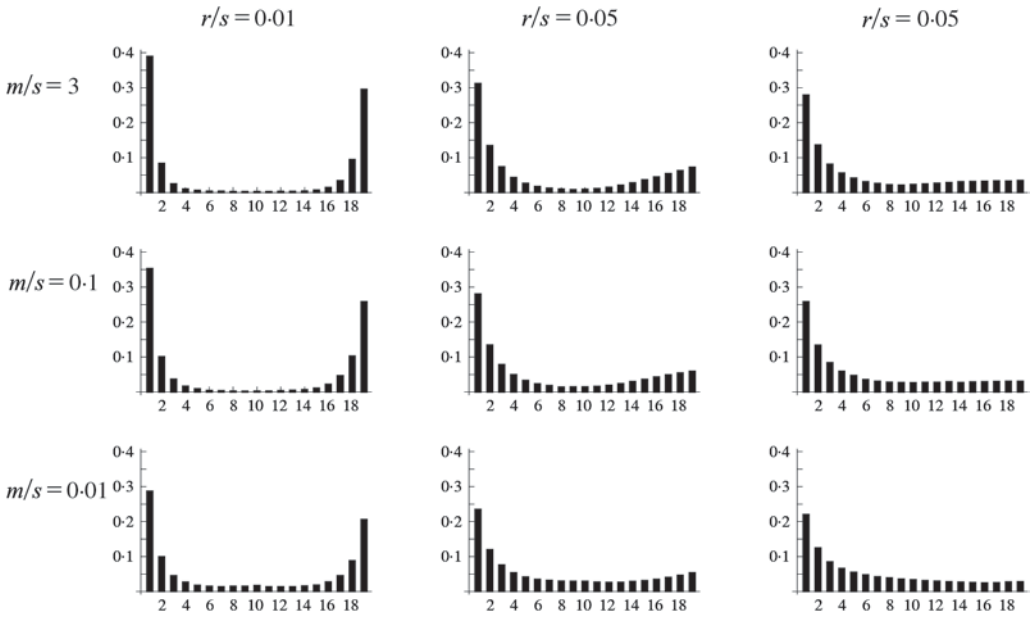

Fig. 1. Site-frequency spectrum as a function of scaled map distance to selective target ($r/s = 0{\cdot}01$, $0{\cdot}05$ and $0{\cdot}09$) and scaled migration rate ($m/s = 3$, $0{\cdot}1$ and $0{\cdot}01$). Sample composition is {(2, 10), (5, 10)}, which means that ten chromosomes are sampled from demes 2 and another 10 from deme 5. Result for each parameter set is based on 100 000 ($K_W = 500$ and $K_S = 200$) replicates. Other parameters: $N = 50\,000$, $s = 0{\cdot}1$, $r_n = 10^{-4}$.

To further examine the change of the relative abundance of intermediate-frequency alleles, Tajima's *D* (Tajima, 1989) was calculated over the 1000 samples of 40 chromosomes with sampling configuration {(2, 20), (4, 20)}. There are 200 neutral loci on a chromosome and, in the results shown in Fig. 2*a*, the sample harbours on average 114·8, 121·1 and 122·6 polymorphic sites after selective sweep when $m = 0{\cdot}3$, $0{\cdot}01$ and $0{\cdot}001$, respectively. Figure 2*a* shows that relative heterozygosity (π) and Tajima's *D* are negatively correlated as expected (Braverman *et al*., 1995). With selection in a panmictic ($m = 0{\cdot}3$) population, Tajima's *D* is mostly negative (94·5% of replicates). However, with significant subdivision ($m = 0{\cdot}001$), less than half of replicates (46%) yield negative *D*. This is partly because the strength of hitchhiking diminished (π increased) with decreasing *m*. For the similar values of π, the mean of Tajima's *D* under population subdivision is still greater than that in panmixia. However, distributions overlap substantially, suggesting that it will be difficult to distinguish sweeps in subdivided population versus panmixia based on Tajima's *D* alone. It was also examined how the scheme of sampling (composition of demes from which chromosomes are sampled) affect Tajima's *D* (Fig. 2*b*). Relative to the effect of changing migration rate, that of sampling scheme was small. In addition to sample compositions used in Fig. 2*b*, more 'unbalanced' configurations ({(2, 36), (5, 4)} and {(2, 29), (4, 8), (5, 3)}) were also tried but resulted in similarly small shifts in Tajima's *D* (data not shown).

(iii) *Linkage disequilibrium*

Another important signature of recent selective sweep is the unique spatial patterning of LD (Kim & Nielsen, 2004; Stephan *et al*., 2006). After the fixation of the beneficial mutation, LD increases between polymorphic sites on the same side, but not on opposite sides, of the B locus. In this study, LD between two neutral loci is measured by the correlation coefficient r^2 (Hill & Robertson, 1968). The mean of r^2 for all possible pairs of polymorphic sites located on the left side (neutral loci from 1 to 100) of a simulated sample, $r^{2(L)}$, was calculated. Similarly, the mean value for the right side, $r^{2(R)}$, was obtained. Figure 3*a* shows that $r^{2(L)}$ is negatively correlated with π, as replicates with fewer hitchhike-breaking recombination events generate higher LD and lower heterozygosity, and increases as *m* decreases. Therefore, population subdivision appears to enhance the effect of a selective sweep in3 increasing LD.

Next, the ω statistic of Kim & Nielsen (2004) that quantifies the spatial patterning of LD specific to selective sweep was calculated. (Basically, $\omega = \{(k_L\, r^{2(L)} + k_R\, r^{2(R)})/(k_L + k_R)\}/r^{2(LR)}$, where k_L (k_R) is the total number of pairs of polymorphic loci on the left (right) side of the chromosome and $r^{2(LR)}$ is the mean value of r^2 for all pairs of loci located on the opposite sides.) This statistic however did not increase with decreasing *m* (Fig. 3*b*) because population subdivision increased $r^{2(LR)}$ as well as $r^{2(L)}$ and $r^{2(R)}$.

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

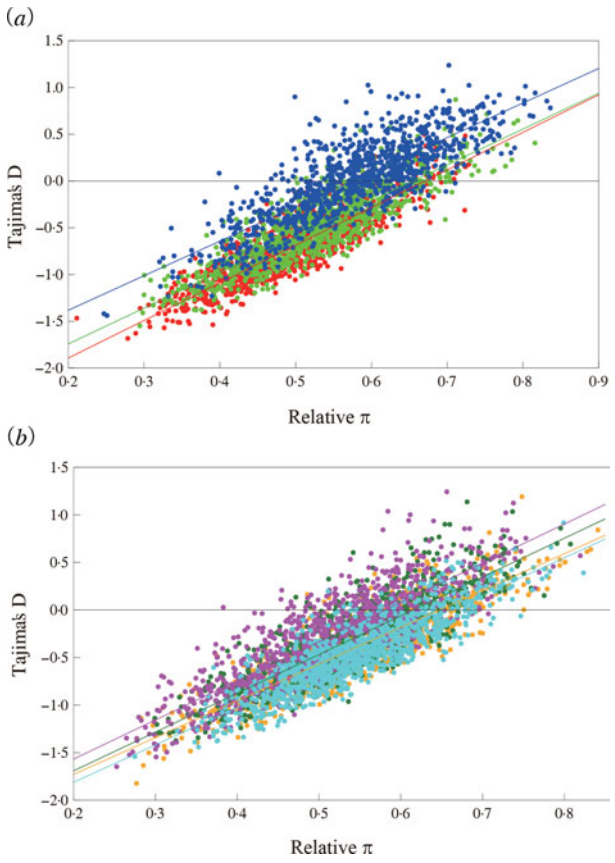

Fig. 2. Joint distribution of Tajima's D and heterozygosity (π) relative to the neutral expectation. (*a*) Full distributions along with the fit of linear model with three values of m are shown for sampling scheme {(2, 20), (5, 20)}: $m=0{\cdot}3$ (red), 0·01 (green) and 0·001 (blue). (*b*) Distributions and linear fit with four different sampling schemes are shown for $m=0{\cdot}005$: {(1, 8), (2, 8), (3, 8), (4, 8), (5, 8)} (orange), {(2, 20), (4, 20)} (green), {(3, 40)} (purple) and {(2, 10), (4, 10), (8, 10), (10, 10)} (cyan). Other parameters: $N=50\,000$, $s=0{\cdot}1$, $r_n=10^{-4}$. Simulation results are based on 1000 replicates ($K_W=200$ and $K_S=5$) for each parameter set.

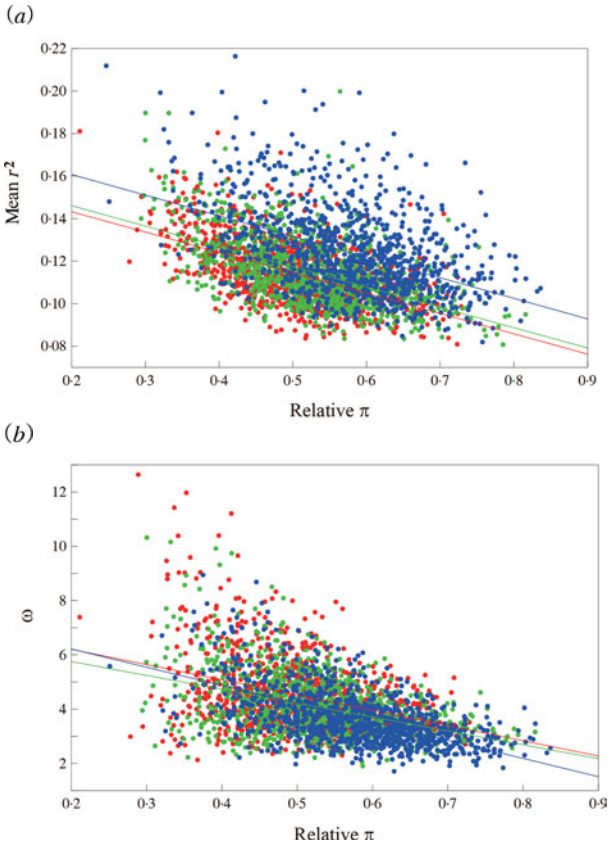

Fig. 3. Joint distribution of the measure of LD ($r^{2(L)}$ (*a*) or ω (*b*)) and relative heterozygosity (π) with sampling scheme {(2, 20), (5, 20)} and $m=0{\cdot}3$ (red), 0·01 (green) and 0·001 (blue). Other parameters are identical to Fig. 2.

(iv) *Genetic differentiation of populations*

As suggested in the previous studies (Slatkin & Wiehe, 1998; Wiehe *et al.*, 2005; Faure *et al.*, 2008; Bierne, 2010), a selective sweep spreading over a subdivided population is expected to increase Wright's F_{ST} at regions flanking the B locus, if the population is initially weakly differentiated. The distributions of F_{ST} observed in the simulated data, with sample configuration {(2, 20), (5, 20)}, are shown in Figure 4. Here, F_{ST} was estimated by $(\pi_T-\pi_W)/\pi_T$, where π_T is the mean pairwise difference between chromosomes over the entire population and π_W is that of chromosomes sampled within a deme (deme 2 or 5) (therefore, identical to K_{ST} of Hudson *et al.* (1992)). The basal distribution of F_{ST} expected for anonymous neutral loci was obtained by simulation run identically to the above but without selection for 400 generations (a slight overestimate of τ). After selective sweeps with $m/s=0{\cdot}1$ and 0·01, 27·1 and 78% of the replicates, respectively, yielded F_{ST} greater than the 97·5 percentile of the corresponding neutral distribution (indicated by dashed lines in Fig. 4). This suggests that one is likely to detect the recent fixation of a positively selected allele using the genomic scan of F_{ST} alone (Bierne, 2010), if the geographical structure of population substantially limits the spread of the allele by selection. With panmixia ($m/s=3$), only 3·5% of replicates yielded F_{ST} greater than the neutral threshold. Results using the 99 percentile as the neutral threshold are also listed in Table 1. Overall, compared with Tajima's D and the measures of LD, F_{ST} is more informative in distinguishing selective sweeps in panmixia vs. population subdivision.

Using deterministic approximations, Slatkin and Wiehe (1998) and Bierne (2010) predicted that the expected F_{ST} after a selective sweep in a subdivided population is greatest at an intermediate distance from the site under selection (i.e. with an intermediate r/s). In agreement with these predictions, simulations with decreasing recombination rates (thus subjecting neutral loci to a stronger hitchhiking effect) yielded declining statistical powers of rejecting neutrality

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

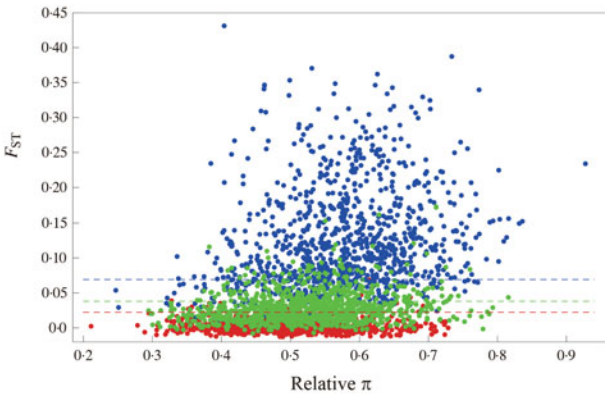

Fig. 4. Joint distribution of F_{ST} and relative heterozygosity (π) after selective sweeps in subdivided populations with sampling scheme {(2, 20), (5, 20)} and $m=0{\cdot}3$ (red), 0·01 (green) and 0·001 (blue). 97·5 percentiles in the distribution of F_{ST} in the corresponding neutral simulations are shown in dashed lines. Other parameters are identical to Fig. 2.

based on F_{ST} alone (Table 1, cases 10–15). However, the degree of decline is rather moderate, while other signatures of selection (e.g. skew of frequency spectrum and increase in LD) are greatly enhanced.

(v) *Gradient of heterozygosity*

A major result in Kim & Maruki (2011) is that the heterozygosity-reducing effect of selection decays as the spatial distance travelled by the beneficial allele increases. Therefore, observing a gradient of after-hitchhiking heterozygosity (i.e. the relative abundance of residual variation at a genomic region where selective sweep already wiped out most polymorphism) across demes may indicate the path of beneficial allele's spread. To assess how reliably such a gradient would be observed in a sample of DNA, which is only one realization of a highly stochastic process, simulations above with sample configuration {(2, 20), (5, 20)} were analysed. The gradient of heterozygosity was quantified by $\Delta\pi=(\pi_5-\pi_2)/(\pi_5+\pi_2)$, where π_2 (π_5) is the expected heterozygosities calculated for sequences sampled in deme 2 (5). The mean of $\Delta\pi$ is positive with $m/s<1$ as expected (Fig. 5*a*). However, the variance of $\Delta\pi$ is very large for all values m examined. Even with $m/s=0{\cdot}01$, 25·4% of replicates yielded $\Delta\pi<0$.

One may test the significance of $\Delta\pi$ by obtaining the null distribution of $\Delta\pi$ using the simulation of selective sweep in an effectively panmictic population. From the simulated dataset with $m/s=3$, the 2·5 and 97·5 percentiles of $\Delta\pi$ were obtained and these values defined cut-offs for rejecting the hypothesis of no spatial structure. Then, with $m/s=0{\cdot}1$ (0·01), 1·0 (3·2)% of replicates yielded significantly negative $\Delta\pi$ and 20·9 (24·5)% of replicates yielded significantly

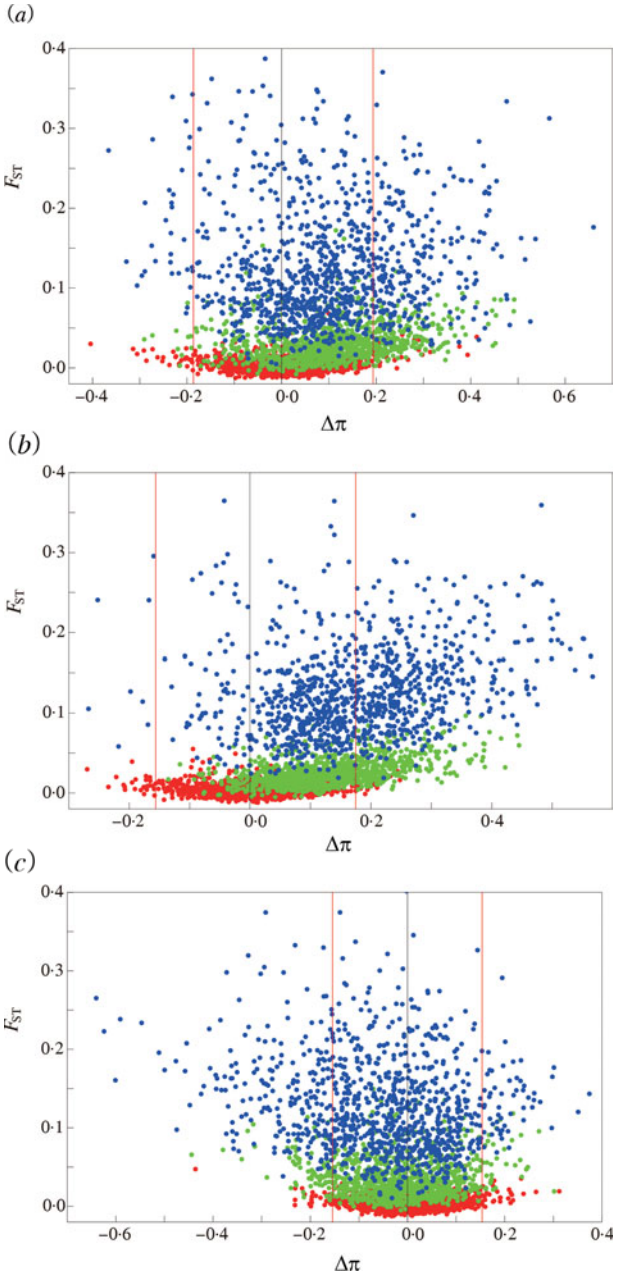

Fig. 5. Joint distribution of F_{ST} and $\delta\pi=(\pi_5-\pi_2)/(\pi_5+\pi_2)$ with sampling scheme {(2, 20), (5, 20)} and $m=0{\cdot}3$ (red), 0·01 (green) and 0·001 (blue). Strength of selection is $s=0{\cdot}1$ for all demes (*a*) or 0·1 for demes 1–3, 9–10 and 0·05 for demes 4–8 (*b*) or 0·05 for demes 1–3, 9–10 and 0·1 for demes 4–8 (*c*). 2·5 and 97·5 percentiles of $\Delta\pi$ in simulations with $m=0{\cdot}3$ are shown by red vertical lines. Other parameters are identical to Fig. 2.

positive values of $\Delta\pi$. Therefore, with less than 25% of chances, the direction of beneficial mutants' spread can be correctly inferred, under scenarios considered here.

This way of inferring the migration path of beneficial mutation further suffers a complication if the strength of directional selection is not uniform across demes. In demes with weaker selection, the effect of

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

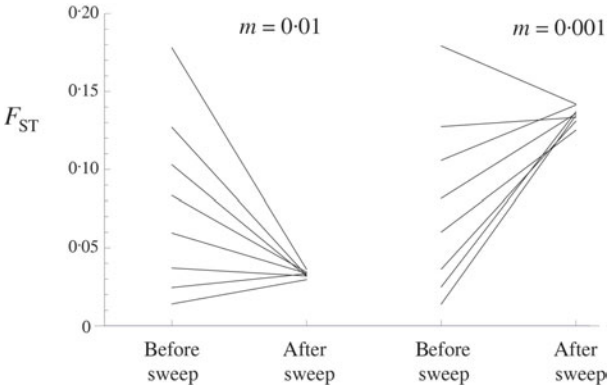

Fig. 6. Change in F_{ST} due to selective sweeps in a subdivided population. Differentiation in the ancestral polymorphism was imposed by the period of isolation ($L_I = 10$, 100, 200, 400, 600, 800, 1000, 1500) at the beginning of simulation. Mean values of F_{ST} before and after selective sweeps (over 1000 replicates) for each set of L_I and m (=0·01 or 0·001) are connected by a solid line. Other parameters: $N = 50\,000$, $s = 0{\cdot}1$, $r_n = 10^{-4}$, sampling scheme = {(2, 20), (5, 20)}.

genetic hitchhiking will also be weaker regardless of its proximity to the geographic origin of the beneficial mutation (Bierne, 2010). To examine this effect, a new set of simulated data was obtained using the same parameters as above but using a reduced selection coefficient $s = 0{\cdot}05$ (instead of 0·1) of the beneficial mutation in demes 4–8. As expected, much positive-biased distribution of $\Delta\pi = (\pi_5 - \pi_2)/(\pi_5 + \pi_2)$ was generated and $\Delta\pi$ greater than the 97·5 percentile of the corresponding null distribution was obtained in the 31·4% (48·5%) of replicates with $m/s = 0{\cdot}1$ (0·01) (Fig. 5*b*). Next, in the reverse scenario, simulation was performed with selection coefficient 0·05 in demes 1, 2, 3, 9 and 10 and 0·1 in other demes. As the hitchhiking effect of the beneficial allele is stronger in deme 5 than in deme 2, the distribution of $\Delta\pi$ shifted to the negative value. Furthermore, 7·3% (21·8%) of replicates with $m/s = 0{\cdot}1$ (0·01) yielded $\Delta\pi$ less than the 2·5 percentile of the corresponding null distribution ($m/s = 3$). This result suggests that an incorrect inference on the direction of beneficial allele's propagation can be obtained if the heterogeneity of selective pressure is not considered.

No strong correlation between $\Delta\pi$ and F_{ST} is observed, particularly with spatially homogeneous selection (Fig. 5*a*). This suggests that an increase in F_{ST} does not imply a difference in heterozygosity between demes, which is however the case in detecting a local selective sweep by F_{ST} (Lewontin & Krakauer, 1973; Beaumont & Balding, 2004), but reflects the varying profiles of haplotypes hitchhiking to the beneficial allele in different demes (Slatkin & Wiehe, 1998). With spatially heterogeneous selection (Figs 5*b* and *c*), F_{ST} shows clearer positive correlation to $|\Delta\pi|$, indicating that the increase of F_{ST} in this case is associated with differential after-sweep heterozygosity due to different strengths of selection across demes.

(vi) *The effect of ancestral population differentiation*

The model of selective sweep investigated so far assumes no genetic differentiation among subpopulations at the start of selective sweep. To examine how robust the pattern of polymorphism generated is to this initial condition, additional simulations were performed using the following procedure that artificially elevates the level of ancestral population differentiation. The forward individual-based simulation, with the same initialization that assigns ancestral numbers to chromosomes, is run for L_I generations ('length of isolation') without any migration of haploids between demes. Then (at time defined again as $t = 0$), a beneficial mutation occurs on a random chromosome in deme 1 and the rest of run is identical to the simulations performed above, including migration of individuals with probability m. At $t = 0$, n chromosomes are sampled (in the same configuration that will be used again at the end of selective sweep) and bi-allelic polymorphism on them is generated as described earlier. From this ancestral polymorphism F_{ST} (the 'before-sweep' value resulting from L_I generations of isolation) is calculated. Varying L_I from 10 to 1500, ancestral differentiation ranging from mean $F_{ST} = 0{\cdot}014$–$0{\cdot}18$ was produced. Figure 6 shows the relationship between this ancestral F_{ST} and the final value of F_{ST} obtained at $t = \tau$ (the 'after-sweep' value). Regardless of ancestral F_{ST}, final F_{ST} is in narrow range that is primarily determined by the migration rate during the selective phase. This result directly confirms earlier studies that a selective sweep in a subdivided population lowers (increases) F_{ST} if population initially exhibits high (low) F_{ST} (Slatkin & Wiehe, 1998; Santiago & Caballero, 2005), and implies that at a certain migration rate a change in F_{ST} would not be observed after a selective sweep. Other patterns of polymorphism (Tajima's D, LD and $\Delta\pi$) were also affected little by ancestral differentiation (Table 1, cases 4, 5, 8 and 9). (A slight increase in Tajima's D is observed, probably because genetic drift during the L_I generations leads to the loss of lineages at low frequencies and thus an excess of alleles at $t = 0$.) In conclusion, the outcome of selective sweep in a subdivided population depends little on the pattern of ancestral polymorphism, justifying the simulation method used above.

4. Discussion

Kim & Maruki (2011) showed that the coalescent process shaping the genealogy in selective sweeps is modified in a subdivided population relative to a

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

panmictic population. This study further examined the effect by computer simulations that generated the stochastic patterns of multi-site polymorphism. The first goal of this study is to examine how the standard signatures of a single selective sweep are affected by population subdivision. Frequency spectrum after sweep showed a clear but minor change in panmictic to subdivided population. With decreasing m, the proportion of extreme-frequency alleles decreased and that of intermediate-frequency alleles increased (Figs 1 and 2). However, this degree of modification may not cause a serious problem in detecting a single selective sweep in a subdivided population by a method based on frequency spectrum expected under the standard model of hitchhiking (e.g. the composite likelihood ratio test of Kim & Stephan, 2002), because the relative proportions of frequency classes did not change. Moreover, distinct shift in frequency spectrum was made only with significantly low values of m relative to s. Unless the strength of selection is very large, small m/s would mean smaller Nm that leads to distinct population structure identifiable by random neutral markers. In such a case, it is less likely that sequences from different demes be pooled into one sample to be analysed. The level of LD after selective sweep was not much affected by population subdivision either (Fig 3*a*). Moreover, the ω statistic, designed for detecting the characteristic spatial patterning of LD, was found to be quite robust to population subdivision (Fig. 3*b*).

Previous studies showed that unique patterns of variation arise in selective sweep in a subdivided population compared with that in a panmictic population (Slatkin & Wiehe, 1998; Bierne, 2010; Kim & Maruki, 2011). Identifying such a pattern in an actual sample of DNA would add further statistical support in detecting selective sweeps or provide information for inferring the demographic context in which a selective sweep occurs. The second goal of this study is to assess how probable it would be to observe such patterns in a finite sample of sequences. As predicted by Slatkin & Wiehe (1998), a selective sweep spreading across demes may bring different haplotypes to high frequency in different demes if the neutral loci are partially linked to the beneficial allele. This leads to an increase of F_{ST} at regions surrounding the locus on selection. Our simulation showed that the probability of obtaining a significant value of F_{ST} (larger than expected under the same population structure but in the absence of selection) can be large if m is much smaller than s (Fig. 4). Therefore, a genomic scan for outliers of F_{ST}, frequently performed to detect the locus of local adaptation (Lewontin & Krakauer, 1973; Beaumont & Balding, 2004), may also detect recent directional selection that drove the fixation of a beneficial mutation in all demes (Bierne, 2010).

As implicitly suggested by Slatkin & Wiehe (1998) and explicitly predicted by Kim & Maruki (2011), the after-sweep expected heterozygosity increases along the migration path of beneficial allele in the stepping-stone model of population subdivision. Therefore, the geographical gradient of heterozygosity may allow us to infer direction of beneficial allele's spread. However, simulations for sequences sampled from two demes (one located closer to the origin of mutation than the other) in this study (Fig. 5) suggests that the expected (mean) difference in heterozygosity between demes is not large enough, relative to the variance of heterozygosity that results from the highly stochastic nature of hitchhiking-affected coalescent-with-recombination process, to allow such an inference. Moreover, the difference in local selective pressure can also lead to a spatial gradient of heterozygosity and thus complicate the inference. The statistical power of inference may increase as sequences are sampled from more than two demes. However, this would not eliminate the problem of the spatial heterogeneity of selection.

This study used individual-based simulation to investigate selective sweeps in subdivided populations. Simulation time to generate bi-allelic polymorphism was greatly reduced by making a proper assumption about ancestral polymorphism at the time of mutation to a beneficial allele. A similar method of combining pre-determined ancestral polymorphism with forward-in-time individual-based simulation to investigate the stochastic nature of selective sweeps was used in Messer & Neher (2012). They used coalescent simulation to generate neutral bi-allelic polymorphism for a large sample of sequences and treated it as the whole-population profile of polymorphism. This approach is superior to the method of this study regarding the correlation of ancestral polymorphism between adjacent sites. However, their method requires to handle a much larger number of loci on a chromosome in order to generate an equivalent number of polymorphic sites in the sample because, as ancestral polymorphism is assigned to the whole population rather than to lineages that are ancestral to the sample, these ancestral lineages at many sites will be monomorphic. Coalescent simulation, used more commonly in population genetic studies than individual-based simulations, could also be used for this study. Indeed, program msms (Ewing & Hermisson, 2010), which is based on the structured coalescent conditional on the trajectory of beneficial allele determined by forward simulation, should be able to simulate polymorphism under selective sweep spreading over structured population. (However, the current version of msms does not provide options – sampling sequences at the time of beneficial allele's fixation while specifying the deme of beneficial allele's origin – that allows the simulation of the exact

https://doi.org/10.1017/S0016672313000062 Published online by Cambridge University Press

scenario considered in this study.) Although coalescent simulation might be better in speed and is free of issues regarding the ancestral polymorphism, it does not allow an easy modification when one wants to add more complexity (about the genetic and demographic details of natural selection) to a model. On the other hand, it is much easier to build and explore complex models using an individual based simulation. Therefore, the approach developed in this study is potentially very useful for a wider community of researchers in their investigation of advanced models.

I would like to thank three anonymous reviewers whose comments greatly improved the manuscript. This study was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (Grant number 2012R1A1A2004932) to Y. K.

5. Declaration of Interest

None.

References

Beaumont, M. A. & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969–980.

Bierne, N. (2010). The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution* **64**, 3254–3272.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796.

Ewing, G. & Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065.

Faure, M. F., David, P., Bonhomme, F. & Bierne, N. (2008). Genetic hitchhiking in a subdivided population of *Mytilus edulis*. *BMC Evolutionary Biology* **8**, 164.

Fay, J. C. & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG: Theoretical and Applied Genetics* **38**, 226–231.

Hudson, R. R., Boos, D. D. & Kaplan, N. L. (1992). A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution* **9**, 138–151.

Kaplan, N. L., Hudson, R. R. & Langley, C. H. (1989). The 'Hitchhiking Effect' revisited. *Genetics* **123**, 887–899.

Kim, Y. & Maruki, T. (2011). Hitchhiking effect of a beneficial mutation spreading in a subdivided population. *Genetics* **189**, 213–226.

Kim, Y. & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513–1524.

Kim, Y. & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777.

Kim, Y. & Stephan, W. (2003). Selective sweeps in the presence of interference among partially linked loci. *Genetics* **164**, 389–398.

Kim, Y. & Wiehe, T. (2009). Simulation of DNA sequence evolution under models of recent directional selection. *Briefings in Bioinformatics* **10**, 84–96.

Lewontin, R. C. & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195.

Maynard Smith, J. & Haigh, J. (1974). The hitch-hiking effect of a favorable gene. *Genetical Research* **23**, 23–35.

Messer, P. W. & Neher, R. A. (2012). Estimating the strength of selective sweeps from deep population diversity data. *Genetics* **191**, 593–605.

Ralph, P. & Coop, G. (2010). Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics* **186**, 647–668.

Santiago, E. & Caballero, A. (2005). Variation after a selective sweep in a subdivided population. *Genetics* **169**, 475–483.

Slatkin, M. & Wiehe, T. (1998). Genetic hitch-hiking in a subdivided population. *Genetical Research* **71**, 155–160.

Spencer, C. C. A. & Coop, G. (2004). SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* **20**, 3673–3675.

Stephan, W., Song, Y. S. & Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**, 2647–2663.

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

Wall, J. D. & Hudson, R. R. (2001). Coalescent simulations and statistical tests of neutrality. *Molecular Biology and Evolution* **18**, 1134–1135.

Wiehe, T., Schmid, K. & Stephan, W. (2005). Selective sweeps in structured populations—empirical evidence and theoretical studies. In *Selective Sweep* (ed. D. Nurminsky) Kluwer Academic/Plenum Publishers, New York (pp. 104–117).