Patterns of neutral diversity under general models of selective sweeps

Graham Coop ‡, Peter Ralph ‡

‡ *Dept. of Evolution and Ecology, and Center for Population Biology, University of California, Davis*

Preprint on ArXiv linked to from gcbias.org

Introduction

There is now good evidence that linked selection plays an important role in shaping patterns of genome-wide diversity. However, models of recurrent sweeps are overly simple as they assume that all sweeps are destined for rapid fixation.

In a large population, only the initial fast behaviour of selected alleles affect the coalescent at partially linked sites. Using this intuition we develop a general model of recurrent partial and soft selective sweeps in a coalescent framework using a coalescent with multiple mergers.

Recurrent sweeps of selected alleles to intermediate frequencies can have a profound effect on levels of diversity but can strongly modify other predictions of the hitchhiking model.

A simple model of a partial sweep

Imagine a selected allele which follows one of the three trajectories shown in Figure 1A and a neutral site a genetic distance r away.

If only partially linked ($r \gg 1/N$ and $r \gg \tau$) the lineages at the neutral site only care about the fast behaviour of the selected allele, and so i out of k will be forced to coalesce at time τ approximately with probability

$$\binom{k}{i} q^i (1-q)^{k-i} \quad \text{where} \quad q \approx x e^{-rt_x} \qquad (1)$$

where q is the probability that a lineage does not escape the sweep. This approximation works well at describing patterns of diversity at sites partially linked loci (e.g. Figure 1B). More generally if the trajectory of our selected allele is $X(t)$ we can use $q(r,X) = r\int_0^\infty e^{-rt} X(t)dt$

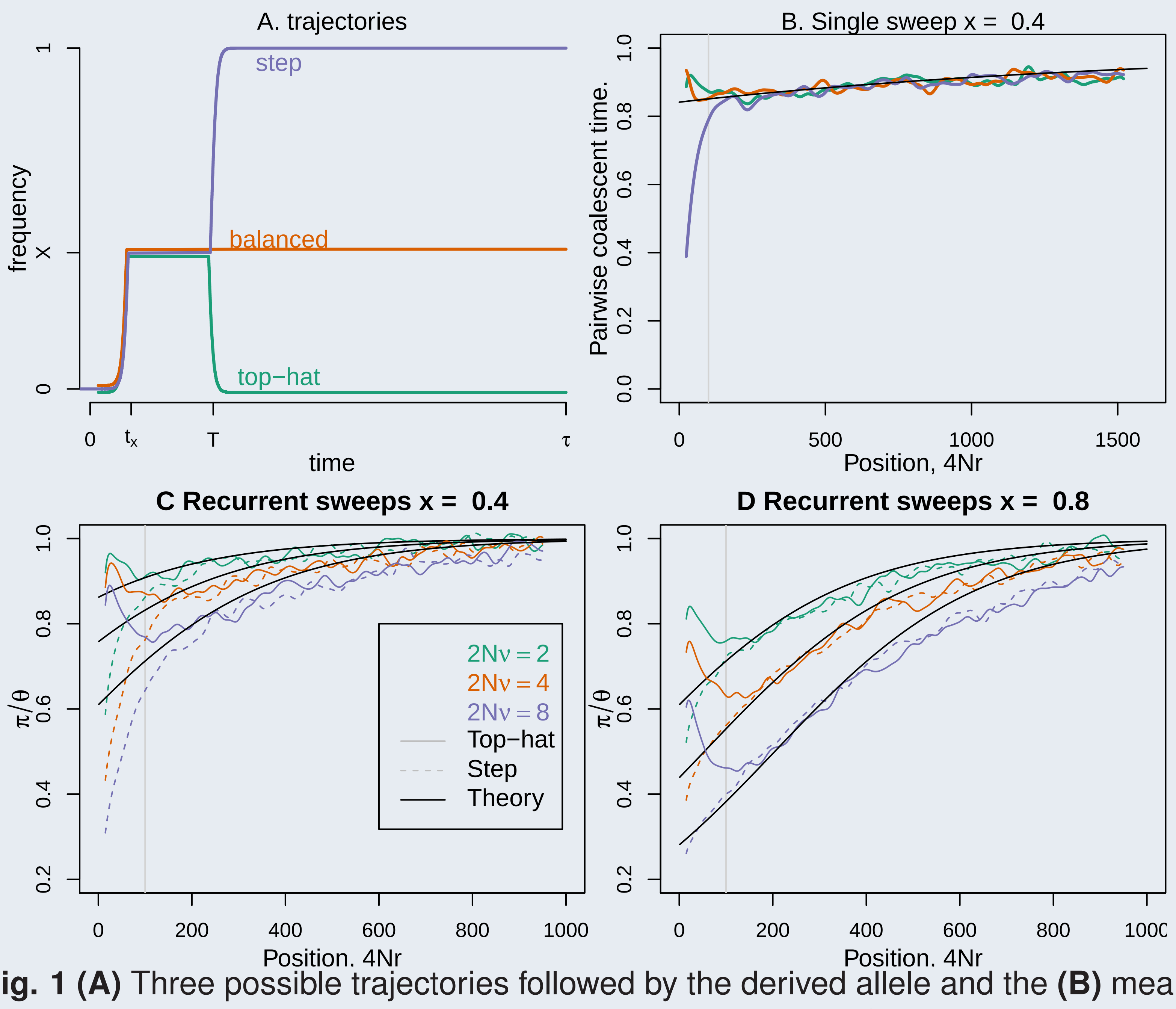

Fig. 1 (A) Three possible trajectories followed by the derived allele and the **(B)** mean pairwise coalescent time against recombination distance away from a selected site that has experienced one of these three sweeps. **(C,D)** the reduction in diversity as a function of recombination distance from a site experiencing recurrent sweeps (either recurrent top-hat trajectories or recurrent step trajectories). The solid black lines give predictions based on our simple binomial model.

A recurrent sweep model Consider a neutral locus partially linked to selected loci at a low rate ν, in a diploid population of size $2N$. Each sweep has its own draw of q, as the recombination distance (r) and trajectory $(X(t))$ may differ across trajectories,

Following from our assumption that each lineage is on the swept background independently with probability q, i out of k lineages, coalesce at rate

$$\lambda_{k,i} = \delta_{i,2}/(2N) + \nu \binom{k}{i} \int_0^1 q^i (1-q)^{k-i} f(q) dq \quad \text{for } 2 \le i \le k. \qquad (2)$$

This multiple mergers approximation works well, see Figure 1C,D, and generalizes the models of recurrent full sweeps [see 3, 2, 1].

Homogeneous partial sweeps

Assuming that sweeps occur homogeneously at rate ν_{BP} along a genome recombining at rate r_{BP} then

$$\lambda_{k,i} = \frac{1}{2N}\binom{k}{2}\delta_{i,2} + \frac{\nu_{BP}}{r_{BP}} J_{k,i}, \quad \text{for } 2 \le i \le k, \qquad (3)$$

where the $J_{k,i}$ are an average over trajectories, and do not depend on ν_{BP} or r_{BP}

$$J_{k,i} = \binom{k}{i} \mathbb{E}_X \left[\int_0^\infty q(r,X)^i (1-q(r,X))^{k-i} dr \right]. \qquad (4)$$

The expected relationship between r_{BP} and polymorphism is:

$$\mathbb{E}[\pi] = \frac{4Nu}{2N\nu_{BP} J_{2,2}/r_{BP} + 1} \qquad (5)$$

which is the same as the full sweep model [3, 5]. Therefore, partial sweeps are just as good an explanation of the relationship between π and r_{BP}, we simply need to turn up the rate of sweeps, e.g. $J_{2,2} = x^2/t_x$ under our simple model. However, for a given reduction in heterozygosity, the skew towards rare alleles is much less when sweeps only reach intermediate or low frequencies (Fig. 2).

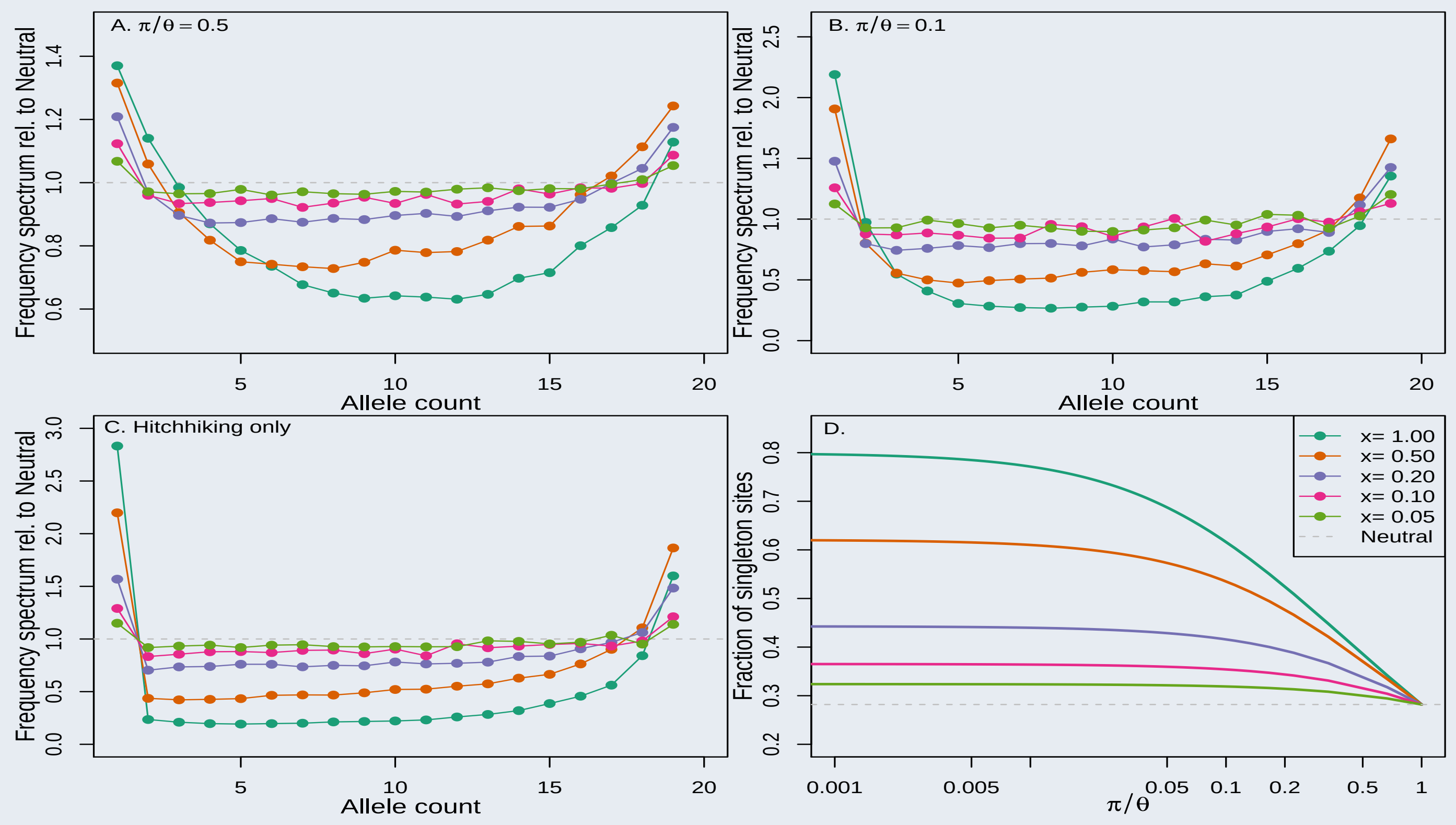

Fig. 2. Properties of the frequency spectrum under a spatially homogeneous model of sweeps across a range of reductions in diversity.

Homogeneous soft-sweeps

Under soft sweep models each sweep forces lineages to coalesce into a set of families, e.g. due to recurrent mutation at the selected site, with a population-scaled mutation rate $4N\rho$ at the selected locus (fig. 3A) [4]. We can then generalize this idea to recurrent soft sweeps occurring homogeneously along the sequence (Fig 3B) to obtain the frequency spectrum for a fixed reduction in heterozygosity (Fig. 3C).

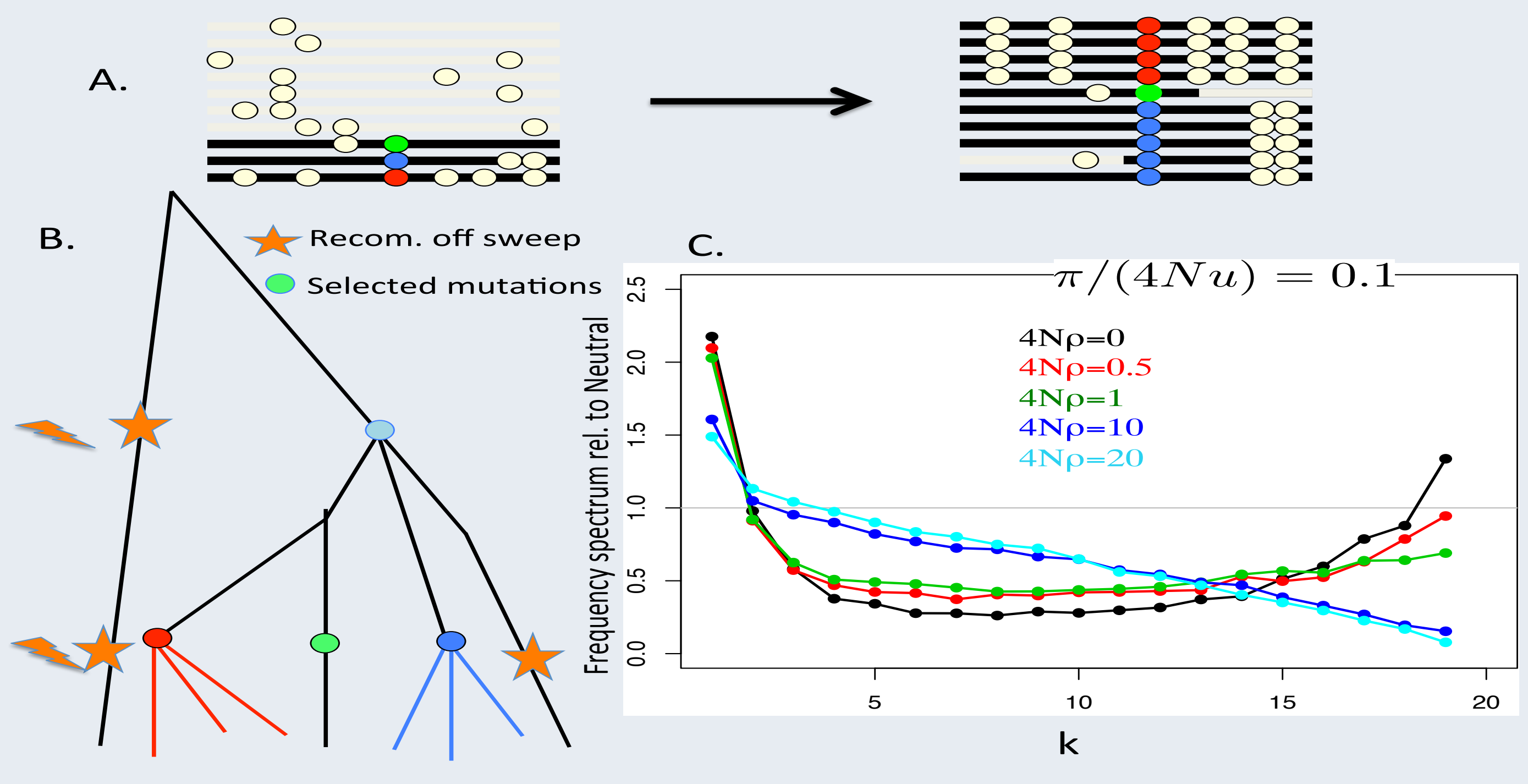

Fig. 3. Recurrent soft sweeps..

References

[1] R. Durrett and J. Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115:1628âĂŞ1657, 2005.

[2] J. H. Gillespie. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics*, 155:909–919, 2000.

[3] N. L. Kaplan, R. R. Hudson, and C. H. Langley. The hitchhiking effect revisited. *Genetics*, 123:887–899, 1989.

[4] P. S. Pennings and J. Hermisson. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.*, 2:e186, 2006.

[5] W. Stephan, T. Wiehe, and M. Lenz. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Popul. Biol.*, 41:237–254, 1992.