J Mol Evol (2003) 57:S190–S200
DOI: 10.1007/s00239-003-0027-y

JOURNAL OF MOLECULAR EVOLUTION

Power of Neutrality Tests to Detect Bottlenecks and Hitchhiking

Frantz Depaulis, Sylvain Mousset, Michel Veuille

Laboratoire d'Ecologie, CNRS UMR 7625-EPHE, Université Pierre-et-Marie Curie, Paris, France

Received: 31 July 2002 / Accepted: 27 October 2002

Abstract. The power of several neutrality tests to reject a simple bottleneck model is examined in a coalescent framework. Several tests are considered including some relying on the frequency spectrum of mutations and some reflecting the linkage disequilibrium structure of the data. We evaluate the effect of the age and of the strength of the bottleneck, and their interaction. We contrast two qualitatively different bottleneck effects depending on their strength. In genealogical terms, during severe bottlenecks, all lineages coalesce leading to a star-like gene genealogy of the sample. Some time after the bottleneck, once new mutations have arisen, they tend to show an excess of rare variants and a slight excess of haplotypes. On the contrary, more moderate bottlenecks allow several lineages to survive the demographic crash, leading to a balanced genealogy with long internal branches. Soon after the event, data tend to show an excess of intermediate frequency variants and a deficit of haplotypes. We show that for moderate sequencing efforts, severe bottlenecks can be detected only after an intermediate time period has allowed for mutations to occur, preferably by frequency spectrum statistics. Moderate bottlenecks can be more easily detected for more recent events, especially using haplotype statistics. Finally, for a single locus, the bottleneck results closely approximate those of a simple hitchhiking model. The main difference concerns the frequency distribution of mutations and haplotypes after moderate perturbations. Hitchhiking increases the number of rare ancestral mutations and leads to a more predominant major haplotype class. Thus, despite a number of common features between the two processes, hitchhiking cannot be strictly modeled by bottlenecks.

Key words: Nucleotide polymorphism — Neutrality tests — Haplotype — Coalescence — Bottleneck — Hitchhiking

Introduction

Our null hypothesis about molecular evolution is expressed in neutral models. They provide predictions about the evolution under mutation drift equilibrium in the absence of systematic effects such as selection or demographic effects. Departures from this model allow us to make inferences about the effects of perturbations on the genetic history of a sample. At the intraspecific level, probably the most simple and commonly used model is that of Wright-Fisher. Besides the neutrality of mutations, it assumes a constant isolated panmictic population with a Poisson distribution of offspring, having reached mutation drift equilibrium. Those assumptions can however usually be relaxed and the effect of the different factors can potentially be explicitly implemented in the model at the corresponding cost of additional parameters to be estimated from the data (see, e.g., Galtier et al. 2000). For natural populations, probably the most unrealistic assumption of this model is that of a constant population size. Natural popula-

Correspondence to: F. Depaulis, Université Pierre-et-Marie Curie, Laboratoire d'Ecologie, CNRS UMR 7625 case 237, 7 quai St Bernard, 75252 Paris Cedex 05, France; *email:* fdepauli@snv.jussieu.fr

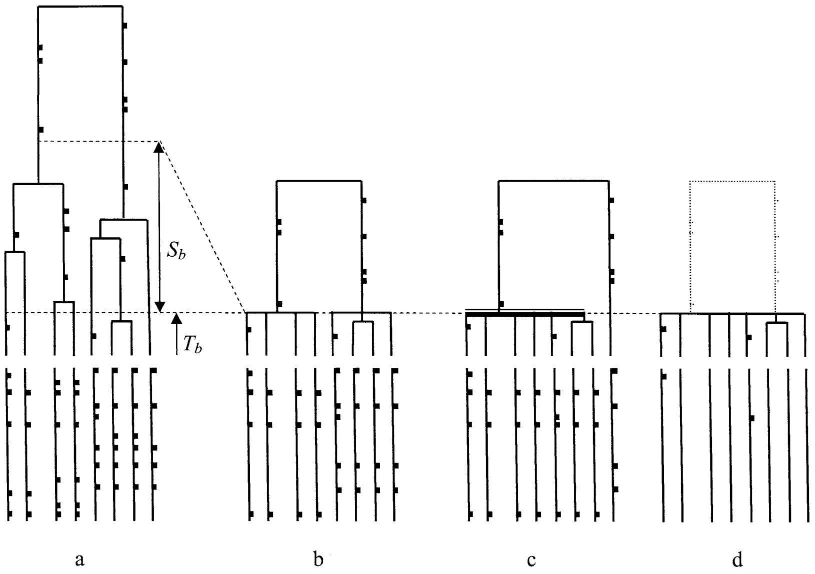

Fig. 1. Outline of the shape of genealogy ($n = 8$) under various models. Mutations are plotted on the tree and on the corresponding sequence sample represented vertically below. T_b, S_b: age and strength of the perturbation (respectively, see text for definitions of parameters), **a** Neutral, constant size type of genealogy ($S = 15$). **b** Genealogy after a moderate bottleneck, two lineages survive the bottleneck stage ($S = 9$) leading to a deficit in the number of haplotypes, an excess of intermediate frequency variants. **c** Genealogy after hitchhiking with recombination of comparable magnitude ($S = 9$). The thick double line represents the lineage carrying the advantageous mutation when it arises. The remaining lineage escaped the sweep through recombination into the advantageous background. The haplotype diversity is more drastically reduced due to the high frequency of the major haplotype. Note also the excess of high frequency derived mutations. **d** After a strong recent bottleneck or a complete selective sweep without recombination, a single lineage survives the selective sweep ($S_b \to \infty$, $S = 2$). Although all mutations are unique and the number of haplotypes is maximal, there is no power to test this, due to the low number of mutations remaining in the sample.

tions usually fluctuate in size or have persistent expansion or contraction trends. It is thus important to be able to detect such departures from the model when they significantly affect the predictions of the model, and to evaluate their impact on the pattern of genetic variation. This involves assessing when the neutral model cannot be applied in its simplest form and when such effects have to be explicitly taken into account in the models.

We describe here bottleneck effects in terms of the modification of the gene genealogy of a sample under a coalescent framework (Hudson 1993; Fig. 1). We also describe the consequence for nucleotide polymorphism datasets as concerns the frequency spectrum of mutations and their association as shown by haplotype distribution, defined by a particular succession of mutations along a sequence or a chromosome. Gene genealogies represent an intuitive way of describing perturbation effects looking backwards in time (below, we use "perturbation" to refer to any event, in particular bottlenecks and hitchhiking, occurring in the genetic history of the sample and leading to a violation of the assumptions of the standard neutral model). As coalescence is a sampling theory, the genealogy outcomes are directly comparable to empirical datasets derived from samples. We contrast two qualitatively different kinds of bottlenecks, depending on their severity. We use "*severe*" to describe bottlenecks leading all lineages to coalesce during the demographic crash, resulting in a reduced genealogy size (the root occurs during the bottleneck stage), with a star-like shape (Fig. 1d). The consequence on a population genetic sample is a reduced level of variation and an excess of rare variants among mutations that arose after the demographic crash. In terms of haplotypes, the resulting samples tend to show an excess of haplotypes with respect to the number of mutations that arose since the event. In contrast, we use "*moderate*" to describe bottlenecks allowing for several lineages to survive the demographic crash (Fig. 1b). Before the bottleneck stage, they maintain (looking backward in time) their large and constant size neutral evolution until they finally coalesce. This leads to a less contracted, balanced genealogy with long internal branches. In a population genetic dataset, the result is a slightly lower level of variation compared to a constant size population, a deficit of rare mutations among those that remain, with most mutations affecting internal branches with a large number of descendants. The number of haplotypes tends to decrease compared to the total number of mutations along the tree.

However, such demographic effects can mimic selective effects. For instance, the effects of bottlenecks on neutral variation on a single locus are similar to selective sweep effects, the hitchhiking effect of an advantageous mutation on linked neutral markers

(Barton 2000). Strong bottlenecks would correspond to hitchhiking with fixation of the advantageous mutant before any recombination had occurred between the selected and the neutral loci, thus removing all variation (Fig. 1d). Moderate bottlenecks predict a pattern similar to hitchhiking with, for example, recombination events occurring between the selected locus and a given marker during the selective stage (Fig. 1b and c, respectively). Alternatively, it may also reflect incomplete sweep, before the fixation of the advantageous mutant, frequency dependent or fluctuating selection (Barton 2000). This analogy between bottlenecks and hitchhiking has two important consequences. From a fundamental point of view, it makes it difficult to distinguish selective from demographic effects. This issue will not be addressed in the present paper but will be dealt with in the Discussion section. The second consequence of this analogy between bottlenecks and selective sweep models is more practical. On a single locus, hitchhiking may be approximated by simple bottleneck models, which can more easily be implemented using computationally demanding methods such as full maximum likelihood methods (Galtier et al. 2000). However, even for a single locus, there are slight differences between bottlenecks and selective sweeps. Consider a lineage present in a genealogy just before the bottleneck. All its descendant lineages after the bottleneck stage could be considered as a family of closely related lineages (e.g., all but the last lineage on Fig. 1c; Barton 1998). This family of lineages concept is similar to that of haplotypes. These families are equivalent to haplotypes if the perturbation event is recent and neither mutation nor recombination has occurred since the perturbation (but only before it). Hitchhiking models tend to predict a distribution of sizes of these families (of frequencies of haplotypes) that is more heterogeneous than with bottlenecks (Barton 1998). In particular, hitchhiking predicts a more frequent major haplotype class compared to bottleneck type models (Barton 1998; Fig. 1b, c).

We address two issues here. First, we evaluate the relative power of different tests when faced with bottleneck effects, as a function of the age and the strength of the bottleneck and of the interaction of these factors. Second, we compare these results with that of a simple selective sweep model. We focus on nucleotide variation and the corresponding infinite site model. This combination type of genetic marker and mutational model tends to show the closest correspondence. Moreover, for a given variation level, this mutational model allows the most powerful inference, due to the assumed absence of homoplasy. We do not pretend or wish to be exhaustive about all existing neutrality tests (which appear to be largely correlated; Fu 1997), nor with the whole space of parameter values, and we do not claim that one test is better than another in all circumstances. We examined the parameter space extensively, but illustrate the results with only a few sets of parameter values representative of the major qualitative trends of the effects involved. We thus show the effect of each factor separately and their interactions whenever relevant. Our goal is to investigate under which circumstances different tests may be applied and how these results may be interpreted.

We show that frequency spectrum statistics tend to be more useful for detecting rather old, severe perturbations, whereas haplotype statistics are primarily useful for detecting moderate, more recent perturbations. Finally hitchhiking tends to show similar results to bottlenecks, but gives more power with haplotype tests relying on the frequencies of haplotypes.

Methods

In this section, we will first rapidly describe the statistics considered. The general simulation framework will then be presented, both for the computation of confidence intervals under the standard model and for the power analyses. Finally we will describe more specifically the simple and generic bottleneck and hitchhiking models used.

Statistics Presented

We present results for two classes of test statistics. The first class of statistics is based on the frequency spectrum of mutations. In fact, the different frequency spectrum statistics seem to use substantially the same source of information, to provide similar results and thus appear largely redundant (Fu 1997). We present only a few of them. Tajima's (1989) D is probably the most classic test. It compares two unbiased estimators of the mutational parameter of the population $\theta = 4N\mu$ (for an autosomal locus, where N is the diploid effective population size and μ the neutral mutation rate). They thus show the same expectations under a neutral model. The first estimator is the average pairwise difference π (average number of mismatches between two sequences), and the second is Watterson's (1975) θ_W estimator based on the number of polymorphic sites S in a sample of size n. The latter is more sensitive to low frequency variants and negative values of Tajima's D reflect an excess of rare variants.

As an example of a statistic that takes into account the polarity of mutations (whether one state is ancestral or derived), we also consider Fu and Li's (1993) D statistic, constructed in a similar way, but comparing θ_W to an estimate of θ based on the number of derived unique mutations (mutations on external branches of a genealogy). Finally, we consider Fay and Wu's (2000) H statistic, which also considers polarized mutations and compares π to θ_H, an estimator that gives more weight to high frequency derived variants. The latter test was however designed primarily to detect characteristic features of selective sweeps and is not expected to be well suited to detecting bottlenecks.

The second class of statistics is made up of a few related statistics based on the linkage disequilibrium structure, especially the distribution of haplotypes (for a review of these statistics and some of their general properties, see Depaulis et al., in press). Their difference and possible redundancy have not been addressed thoroughly before and we will thus consider several of them. Note that as any measure of linkage disequilibrium, these statistics are not independent of the frequencies of mutation and thus on the statistics considered in the previous class (Nordborg and Tavaré 2002).

The haplotype partition test *HP* tests for the occurrence of a subset of sequences with low variation, a major "haplotype class" (Hudson et al. 1994). For simplicity, we consider a restricted version of the *HP* test, which simply tests for the frequency of the major haplotype. The haplotype number statistic *K* was considered independently, with slightly different approaches, by several authors (Strobeck 1987; Fu 1996, 1997; Depaulis and Veuille 1998; Andolfatto et al. 1999). We use it as in the procedure of Depaulis and Veuille (1998), where it is conditioned on *S*, the number of segregating sites and, potentially can be used as a two-tailed test. The haplotype diversity *H* is a related test, taking into account the frequencies of haplotypes (Depaulis and Veuille 1998). It considers haplotype diversity $H = 1 - \sum p_i^2$ where p_i is the relative frequency of haplotype *i* in the sample. The three preceding tests are correlated: for a given number of segregating sites, a low number of haplotypes tends to be associated with low haplotype diversity and possibly a large subset with reduced variation. We will sum up such patterns by the term "strong haplotype structure." Note however that for a given number of haplotypes, *K*, their relative contributions (frequency), especially that of the major haplotype, affect both *H* and *HP*. Finally we consider a test which reflects other aspects of linkage disequilibrium structure: the Z_{nS} test, based on the average pairwise allelic correlation coefficient (Kelly 1997).

Another possible class of statistics considers the distribution of pairwise differences. However, there are not very powerful statistics as they depend largely on one highly stochastic coalescent event leading to the root of the genealogy (Felsenstein 1992; Rozas, personal communication). They will not be considered here.

In what follows, the tests are evaluated separately for each direction of departure (excess or deficit), but for the sake of clarity only curves showing power above the chosen nominal rejection level (2.5% on each side) on a given figure are presented.

Simulations

We use coalescent simulations, as described in (Hudson 1993), to compute the confidence intervals of the tests and estimate their power under two scenarios—bottlenecks and hitchhiking. For simplicity and without loss of generality, we consider the case of an autosomal marker in a diploid population, and express all time-scaled parameter values in units of *4N* generations, where *N* is the current population size after the bottleneck. The effects of the sampling strategy (sample size, mutational parameter value) have been studied elsewhere (Wall 1999; Depaulis et al., in press). Briefly, increasing information, either in terms of sample size or mutation rate, improves the power of the tests, up to a point where the number of haplotypes saturates and haplotype tests become of little use. This could be reached for an extreme mutation rate over sample size ratio, but such relative ranges of values look unrealistic for most practical cases and this difficulty can be overcome by sliding window approaches (Depaulis et al., in press). We thus fix sample size and variation level at realistic values ($n = 20$ and $\theta = 10$, respectively), which provide reasonable information with most statistics. Our purpose is to compare the results of the tests for a given sequencing effort. Intragenic recombination effects have also been studied elsewhere and have been shown to be substantial in some instances (Wall 1999; Depaulis et al., in press). This effect was examined in our survey and appears very similar to that under hitchhiking, though these results are not reported here to avoid redundancy (Depaulis et al., in press). We restrict the presentation of our results to models without intragenic recombination, and only mention how the results are affected when appropriate (see also the Discussion section). The algorithm was computed independently by the two first authors and their results were cross-checked.

For the power analysis, our simulation procedure was similar to that of Wall and Hudson (2001), who used it for another kind of application, the robustness of conditioning neutral simulations on *S* compared to conditioning on θ. We use standard coalescent simulations conditional on θ (Hudson 1993) for the simulations under the alternative hypothesis models (bottlenecks and selective sweeps). However, we tested the outcome of these power simulations conditioning the tests on the resulting number of segregating sites *S*. [In the absence of intragenic recombination, this is where the haplotype distribution departs from that of Ewens (1972) under the infinite allele model, which conditions on θ.] This approach is probably more relevant in most cases and has been justified elsewhere (Hudson 1993; Depaulis et al. 2001; Wall and Hudson 2001; Przeworski 2002; Depaulis et al., in press). It simply reflects the common case where a population has a given mutational parameter (θ) value, which is however unknown when applying the tests to data (it is highly dependent on the locus considered and the sampling scheme). Only the number of remaining mutations *S* can be readily observed in the data. Our procedure is computationally more demanding than conditioning on either θ or *S* for all simulations, as, for a given θ, a large variety of *S* values are obtained by simulations and the confidence interval has to be estimated for all of them. We thus used only 5,000 simulations to compute the confidence interval for each possible *S* value. As there is no systematic bias, the potential resulting imprecision should be partly compensated by the large number of *S* values considered for the large number of simulations run for the power analyses (100,000 for each set of parameter values): imprecision in one direction in the computation of the bounds for one *S* value is partially compensated by that in the opposite direction for another *S* value. That is, there is an averaging effect, thereby reducing the variances. Such a realistic procedure leads to an overall reduction in the power of the tests (Depaulis et al., in press), which should be biologically relevant. The reason for this is that we do not use the information contained in the reduction of variation compared to the level expected in the absence of a bottleneck. This reflects that the expected level of variation is generally unknown.

Bottleneck Model

A wide variety of models could be used to describe bottleneck or population expansion scenarios, all of which predict similar effects on haplotype distribution and frequency spectrum, merely characterized by a deficit or an excess of rare variants and of haplotypes, depending on the strength of the effect (Fu 1996, 1997). Here we consider a simple generic model as a crude way of looking at the bottleneck class of models (Galtier et al. 2000; Fig. 1). The advantage of this procedure is that it models bottlenecks with only two parameters: T_b, the age of the bottleneck, and S_b, a compound parameter which describes the strength of the bottleneck. In reality, the strength of the bottleneck depends both on the magnitude of the reduction in population size and on the duration of the bottleneck. In coalescent terms, S_b is the corresponding time that would lead to the same amount of common ancestry events if the size of the population was not reduced (Fig. 1a, b). The downside of the simplicity of this model is the assumption that no mutation occurs during the bottleneck stage. This is equivalent to assuming that the bottleneck does not last long. In any case, what happens during the bottleneck stage probably depends on the details of the model. In most datasets, it is very likely that little power exists to distinguish between different bottleneck scenarios. However, other prior information such as demographic records may be used for this purpose, though applying a test in such a situation seems unnecessary. Our aim is to detect the signature common to any bottleneck scenario, whatever the details of the model may be, e.g., exponential or logistic growth. However, mutations occurring during the bottleneck may contribute noise. In fact, their effect is very similar to that under hitchhiking with various values of the selection coefficient s for a given c/s ratio, though these results are not reported here to avoid redundancy (Depaulis et al., in press). The duration of the selective stage depends on the s value. We show that this variation in s can be approximated by an age effect (scaling the T_b parameter value): the relevant age is that of the beginning of the increase of the advantageous allele frequency (or of the population size expansion), when most common ancestry events occur (Depaulis et al., in press). Note that our approach is similar to that of Fay and Wu (1999), summing-up the bottleneck strength by the ratio of the duration over the population size reduction fraction, which should be a valid approximation if mutation can be neglected during the bottleneck stage. Our primary interest here is how many lineages survive the bottleneck stage (have not coalesced during the bottleneck) and what is their representation (their frequency distribution) after the bottleneck. This model provides a straightforward algorithm for the simulations: the coalescent is simulated as in the standard procedure, but neither mutation nor recombination are allowed to occur during the bottleneck stage (between T_b and $T_b + S_b$).

Hitchhiking Model

We used a classic hitchhiking model following a procedure similar to that of Braverman et al. (1995). Readers interested in a more precise description of the procedure are referred to Fay and Wu (2000) and Depaulis et al. (in press). Briefly, we assume a single hitchhiking event of a given age during the history of the genealogy. This procedure may match typical experimental situations when the studied locus is a candidate for having been affected by a selective event in its history and makes few assumptions about the homogeneity and frequency of selection. It is also more intuitive to understand the effects of a single event of known age rather than a potential superposition of several events of various ages. The selective sweep model uses a deterministic approximation to model the change in the frequency x of the advantageous allele between a low frequency ε (when the advantageous mutation is assumed to have occurred) and a point close to fixation ($x = 1 - \varepsilon$). This approximation tends to underestimate the hitchhiking effect. In a population of finite size, among advantageous mutations, the rare mutations that go to fixation tend to increase in frequency more rapidly than others, in their early stages. Fixation is also reached more rapidly at the end of the selective stage. This issue is discussed elsewhere and is expected to have little effect on the results (Barton 2000; Przeworski 2002; Depaulis et al., in press). During the selective stage, the population is composed of two types of genetic backgrounds, neutral, and advantageous. Coalescence can only occur within a background, and its rate depends on the representation of the background in the population. Another type of event affecting the history of the sample is genetic exchange between the selected site and the marker, occurring with a rate $C = 4Nc$ (per $4N$ generations). This should not be confused with possible intragenic recombination occurring at a different scale and showing additional effects on linkage disequilibrium statistics. The selective stage is divided into many (1000) time-steps in order to determine when such events take place. We use a strong selection parameter value $\alpha = 4Ns = 10{,}000$ (s, the selection coefficient, would correspond to a value of about 0.0025 in *Drosophila*) leading to a short selective stage, but during which recombination can still occur with correspondingly high rate of genetic exchange C. The ratio c over s primarily determines the strength of the hitchhiking effect (Stephan et al. 1992). It reflects the strength of the bottleneck S_b in the previous model. One reason for assuming strong selection is that the deterministic approximation is more appropriate for such cases. Another reason to use this high α value and scale C accordingly is that, as with bottlenecks, what happens during the sweep stage should depend on the details of the model. We thus prefer that this selective stage should not make a large contribution to the accumulation of variation and thus should be brief. There are indeed a wide variety of models that could be used depending on the selection regime (dominance level, frequency dependence, density dependence, fluctuating selection; Barton 2000). It is not clear to what extent the results would differ and how selection generally proceeds in nature, but the frequency trajectory of the advantageous allele should have little effect as long as it goes to fixation rapidly (high α values). The other reason relates to comparisons with the bottleneck model. To reflect the artificial absence of mutation assumed during the bottleneck stage in the previous model (this assumption is not necessary in most bottleneck models), we artificially prevent mutations from occurring during the selective stage, which is a more realistic assumption for a short selective stage. Comparison between the two models requires a rescaling. For a given C value, we use the equivalent S_b ($Eq(S_b)$), which would lead to the same expected reduction of diversity (π/π_0). This strength is inversely related to the rate of genetic exchange between the selected locus and the marker (Stephan et al. 1992). The expected reduction of diversity can be easily derived analytically under the bottleneck model. It equals the reduction of pairwise coalescent times:

$$\frac{E(\pi)}{E(\pi_0)} = 1 - e^{-2T_b}(1 - e^{-2S_b}) \tag{1}$$

An equivalent approximation has been derived under a deterministic selective sweep model immediately after the sweep (Eq. 14d of Stephan et al. 1992):

$$\frac{E(\pi)}{E(\pi_0)} = 1 - \varepsilon^{2c/s} \tag{2}$$

Hence an expression to compute the equivalent S_b under the hitchhiking model (setting T_b to 0):

$$Eq(S_b) = -\tfrac{1}{2}\ln(1 - \varepsilon^{2c/s}) \tag{3}$$

Results

We will first describe the effects of severe and moderate bottlenecks as a function of their age, then their strength effect for very recent and older bottlenecks. Finally, we will compare the bottleneck results to those of the hitchhiking model.

Severe Versus Moderate Bottlenecks and Age Effects

Figure 2 shows the effect of a severe bottleneck ($S_b \to \infty$) on the power of haplotype tests as a function of the age of the bottleneck T_b. This case corresponds to all lineages coalescing during the bottleneck, leading to a star-like genealogy with the age of the root corresponding to that of the bottleneck (T_b; Fig. 1d). As predicted, for a given number of polymorphic sites, a severe bottleneck leads to an excess of rare frequency variants (negative Tajima's D and Fu and Li's D) and an excess of haplotypes, of haplotype diversity, and a major haplotype that is underrepresented. The time effect shows a mode for intermediate age for all statistics: if the event is too recent, variation has not recovered enough to detect any depar-

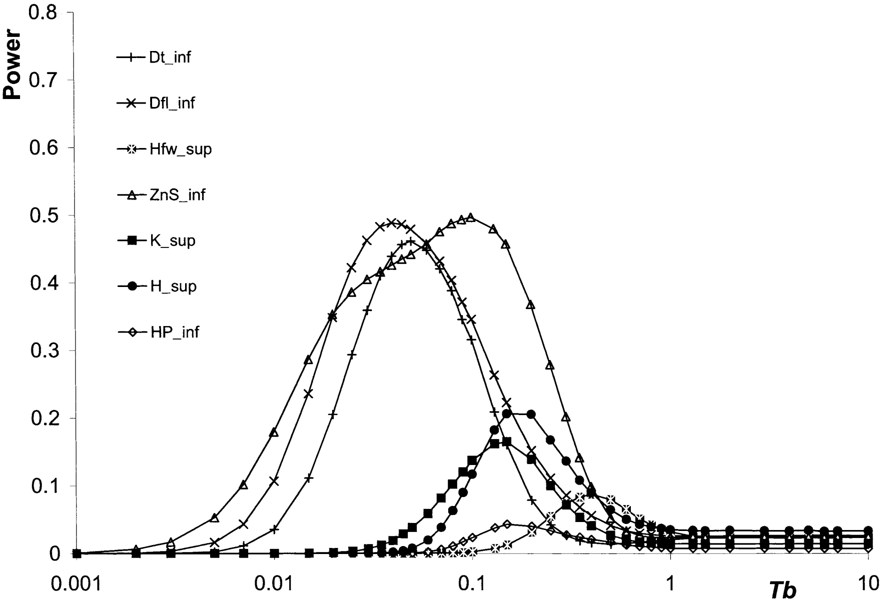

Fig. 2. Power of several neutrality tests as a function of the age (*Tb* in $4N$ generations) of a severe (S_b = 1000) bottleneck (n = 20, θ = 10, Nr = 0). Tests: K, haplotype number; H, haplotype diversity, HP, frequency of the major haplotype; Dt, Tajima's *D*; Dfl, Fu and Li's *D*; Hfw, Fay and Wu's *H*; ZnS, average pairwise allelic correlation. Empty symbols: _inf, significant deficit of the statistic. Filled symbols: _sup, significant excess.

ture from neutrality, there is some power for intermediate age (0.2–2 N generations; note that the x axis is expressed in units of $4N$ generations); but the power drops as the event gets older and the genealogy becomes more neutral (hence the power value close to the nominal chosen rejection level of 2.5%). For such events frequency spectrum statistics (Tajima's *D* and Fu and Li's *D*) and to a lesser extent Z_{nS} tend to show higher peaks of power for more recent events (0.04–2 N generations). The haplotype tests and Fay and Wu's *H* (upperbound) show much less power and show their peak for a somewhat older age (around 0.6 N generations). Of course, power should recover more rapidly with higher θ and n values. What matters is the number of mutations arisen since the bottleneck, for which the expected value can be approximated by $n\theta T_b$, as long as the genealogy is star-like (low nT_b and high S_b values). This product must be substantially higher than 1 to get reasonable power (for high variation, e.g., θ = 50 and n = 50, this may still require more than 10^{-2} N generations; i.e., roughly 10,000 *Drosophila* generations). In this direction, haplotype tests and Z_{nS} are not conservative with respect to intragenic recombination and are probably of little use, unless they can be applied using a conservatively high value of the recombination rate or on non-recombining genetic systems (Wall 1999; Depaulis et al., in press). In addition, intragenic recombination substantially reduces the power of all tests in an effect similar to that described by Wall (1999) (results not shown). In summary, such severe bottleneck effects are difficult to detect simply because there is no variation left shortly after the bottleneck. This counterintuitive effect arises from the fact that we condition the test on the S value remaining after the bottleneck, without taking into account the reduction of variation (assuming that the expected variation level in the absence of bottleneck is unknown). Only moderately old events can be detected, preferably using frequency spectrum statistics, especially if genotyping facilities and the variation level in the sample are both limited.

Moderate bottleneck effects, which do not remove all the preexisting variation but affect its distribution (Fig. 1b), may be easier to detect. Figure 3 shows the effect of the age of a more moderate bottleneck (S_b = 0.25, i.e., N generations), with several lineages generally surviving the bottleneck stage. The high power values obtained for recent bottlenecks correspond to balanced genealogies due to the survival of several lineages during the bottleneck stage (Fig. 1b). Tests thus show some power in the direction of an excess of haplotype structure (a deficit of haplotypes) and a deficit of rare variants (positive frequency spectrum statistics). Note that in this direction, all tests are conservative with respect to intragenic recombination. The power of the tests reaches higher values due to the substantial variation remaining in the sample. Haplotype tests and Z_{nS} tend to show the highest power for this range of parameter values. But the power drops more rapidly with the age of the bottleneck than under the severe bottleneck case and tests show virtually no power after $0.4N$ generations (Fig. 3), in agreement with the results of Przeworski (2002). Newly arisen mutations since the bottleneck constitute the essence of the signal under a severe bottleneck (preexisting variants have all disappeared). On the contrary, after a moderate bottleneck, the signal arises only through sorting of preexisting alleles during the bottleneck and mutations that arose since this event tend to obscure its signal. Fu and Li's *D* shows substantial power, but

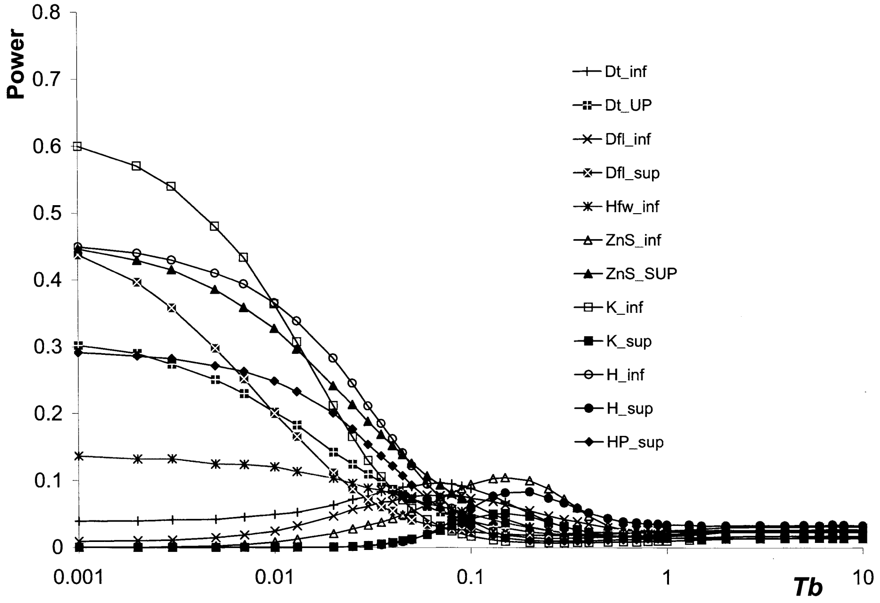

Fig. 3. Power of several neutrality tests as a function of the age of a moderate bottleneck ($S_b = 0.25$ in $4N$ generations). Other parameter values and symbols are as in Fig. 2.

this disappears even more rapidly with the age of the event. In this direction, all statistics are conservative with respect to intragenic recombination. As to the power of the statistics, the effect of intragenic recombination is low for such recent moderate perturbations and realistic recombination rates (results not shown and indistinguishable from Fig. 3). This recombination effect is very similar to that described under a hitchhiking model (Depaulis et al., in press). An intuitive reason for this is that whatever the number of haplotypes and however much recombination has occurred between them before the perturbation, the few haplotypes that survive the perturbation stage will all increase in frequency and will have little time to recombine between each other afterwards. The power does not decrease more rapidly with the age of the event unless the recombination rate is of higher order of magnitude than the mutation rate, which seems to be a rare case in biological systems. Interestingly, for an intermediate range of bottleneck age values (0.1–2 N generations; Fig. 3), the variance of the statistics is drastically increased and the tests show some (low) power in both directions: due to stochasticity, some genealogies show several lineages surviving the bottleneck while others do not, leading to very different values of the statistics (increased variance). This "two-tail" effect is stronger for the intermediate range of parameter values (S_b ranging from 0.25 to 1,000; results not shown). For older events, looking backward in time, many lineages have coalesced before reaching the bottleneck stage and remaining ones may well coalesce during this stage. Thus, a single lineage generally survives the bottleneck (the bottleneck is not strong, but quite old). A substantial amount of time has also elapsed since the event, allowing new mutations to occur on external branches. As a consequence, the tests show some (low) power only in the direction of a deficit of haplotype structure, an excess of rare variants, reflecting a pattern similar to the severe, moderately recent case shown on Fig. 2. The power completely disappears (2.5% constant size level) when the age of the bottleneck reaches the expected age of the most recent common ancestor of a neutral genealogy (i.e., close to $4N$ generations).

Recent Versus Older Bottlenecks and Strength Effects

Figure 4 shows the power of the tests faced with a very recent bottleneck ($0.004N$ generations, about 400 years for *Drosophila*) as a function of its strength. For severe effects, on the right ($S_b > 8N$), there is no power to detect bottlenecks as there is no variation left after the event and the variation did not yet have time to recover (reflecting the left side of Fig. 2). On the contrary, on the left of Fig. 4, the bottleneck effect is too weak to perturb significantly the genealogy, hence the 2.5% neutral constant-size rejection level. The peak obtained for intermediate bottleneck strength corresponds to the excess of haplotype structure obtained on the left of Fig. 3. In this case, the observed departure is in the direction of an excess of haplotype structure (e.g., deficit of K), an excess of Z_{nS} of Fu and Li's D and Tajima's D. For such sets of parameter values, haplotype tests, especially K, show the highest power. Interestingly, Fay and Wu's H shows nonnegligible power (up to 15%) in the direction of a deficit of the statistics when faced with bottlenecks (which was not the original target of this statistic).

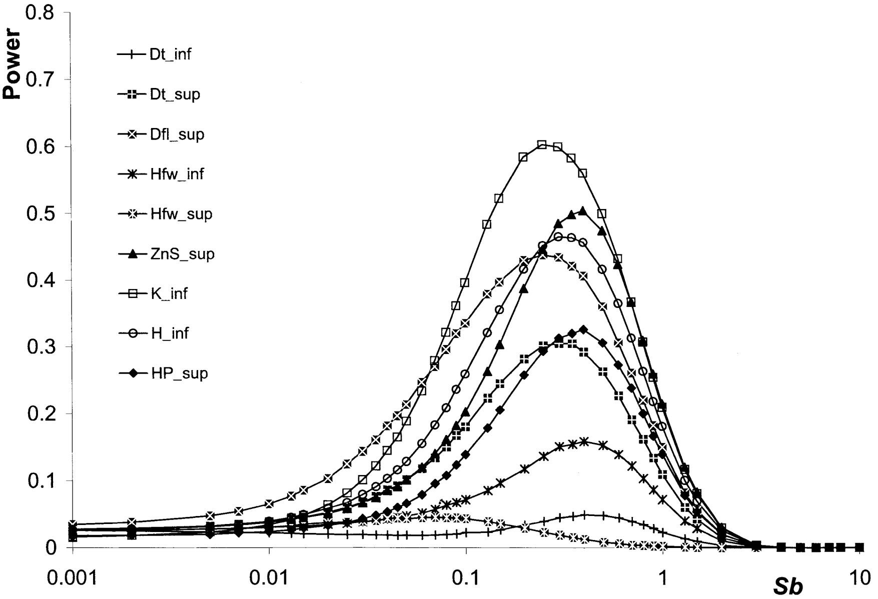

Fig. 4. Power of several neutrality tests as a function of the strength S_b of a recent ($T_b = 0.001$) bottleneck. Other parameter values and symbols are as in Fig. 2.

Figure 5 shows the strength effect on intermediately old events ($0.4N$ generations). The moderate effects are obscured by mutations, which have had time to occur since the bottleneck, and such effects cannot be detected. Only stronger effects ($S_b > N$) can be detected in the opposite direction, reflecting the star-like pattern as in Fig. 2, with primarily a deficit of frequency spectrum statistics of Z_{nS} and a slight deficit of haplotype structure (excess of haplotypes).

Selective Sweeps or Bottlenecks: Differences

To evaluate the limits of the bottleneck model as a proxy for selective sweeps, we compared the previous results to those of selective sweep models. The two models are very similar under severe perturbations (curves essentially superimposed for Fig. 2; results not shown), with both kinds of models leading to star-like trees in such instances (Fig. 1d; Fu 1997). The two models also show very similar age effects, regardless of the strength of the perturbation (Depaulis et al., in press; results not shown and virtually identical to Fig. 5). Of more interest is the case of moderately recent effects (Fig. 1b versus 1c). The difference between the two models lies in the resulting distribution of family sizes of lineages and the corresponding haplotype frequency distribution. Hitchhiking should reduce *H* and increase *HP* (two statistics highly affected by high frequency haplotypes) more dramatically than under a bottleneck. With the bottleneck procedure, the power of *H* and *HP* tests is thus likely to be underestimated compared to hitchhiking types of effects. We illustrate this difference below.

Figure 6 shows the effect of the strength of a recent hitchhiking event ($T_b = 0.001$), reflecting the effect of the genetic distance from the selected site. The results are comparable with bottleneck effects (Fig. 6 versus Fig. 4), except that tests relying on haplotype frequencies (*H* and *HP*) tend to be more powerful under the hitchhiking model. For such sets of parameter values, simply looking at the frequency of the major haplotype (*HP*) seems to provide most of the signature of the selective sweep. However, this is no longer the case for larger sample sizes. Also, tests tend to show most power for stronger perturbations under the hitchhiking model (peaks slightly shifted to the right). This does not seem to be due to the approximation used in Equation 2, as replacing it by the average reduction of diversity across the hitchhiking simulations led to similar results (results not shown). It seems however to reflect rescaling: we fit the first diversity moment under the two models, but higher order moments may be quite different and the power largely depends on those moments. An alternative rescaling based on the reduction in the number of segregating sites did not show such an effect (results not shown). For such moderate perturbation, Tajima's *D* shows power in opposite directions depending on the type of perturbations: *D* is positive with bottlenecks and negative with partial sweeps, leading to an excess of rare variants, partly due to rare ancestral ones (Fig. 1c; Fay and Wu 2000). Tajima's *D* is thus expected to show a departure for a deficit of the statistic, whatever the strength of the hitchhiking (the relative value of the genetic distance from the selected locus and the selection coefficient), but does not allow us to distinguish rather old and strong (closely linked) from recent and moderate (loosely linked) selective sweeps, in contrast with haplotype

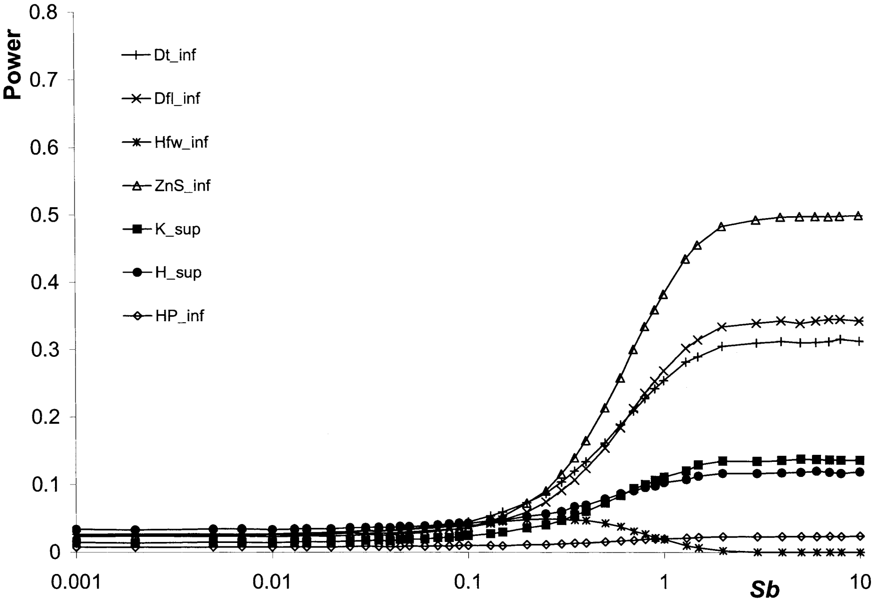

Fig. 5. Power of several neutrality tests as a function of the strength S_b of a moderately old ($T_b = 0.1$) bottleneck. Other parameter values and symbols are as in Fig. 2.

tests. Fu and Li's *D* shows virtually no power for such a set of parameter values. We have studied the properties of neutrality tests faced with this hitchhiking model more extensively elsewhere (Depaulis et al., in press).

Discussion

In summary, moderately old severe bottlenecks can best be detected by statistics relying on the frequency spectrum of mutations. On the contrary, haplotype tests are mainly useful to detect recent and more moderate bottlenecks. Since in practice researchers may not have a precise alternative hypothesis with regard to the severity and the age of the perturbation, it is advisable to combine both types of statistics (with some caution about the multiple testing effects). The present survey should provide help in interpreting the results of the tests. Hitchhiking models produce very similar results, with some slight differences in the frequency spectrum of mutations and haplotypes. In particular, moderate hitchhiking tends to show a more predominant major haplotype class. Thus, hitchhiking cannot be strictly modeled by bottlenecks, though the two processes share a number of common features (compare the present results with that of Depaulis et al., in press). We discuss some limitations of these results and possible extensions below.

The results we illustrated did not involve intragenic recombination (but did involve recombination between the selected site and the marker, on a different scale). There are robustness and power issues regarding this factor. Recombination does not affect the expectation of statistics based on the frequency spectrum of mutations: considering a single polymorphic site, the distribution of the underlying tree and that of the frequency of the mutation at this site are the same whether there is recombination or not. When averaging across sites, the mean is unaffected, but the distribution of such statistics is contracted if there is intragenic recombination as a dataset tends to represent an average across several partially correlated genealogies. As a consequence, such tests, implemented assuming no recombination, are conservative. As recombination tends to break linkage disequilibrium structure, statistics describing such structures are conservative only in the direction of an excess of linkage disequilibrium (excess of Z_{nS}) of enhanced haplotype structure (deficit of *K*, *H*, excess of *HP*; Wall 1999). If there is intragenic recombination, tests may be applied assuming no recombination on the conservative direction for linkage disequilibrium statistics. This is anyway the direction where such tests can be of any use from a power perspective (Figs. 3, 4, 6), and the power is little reduced by recombination for such recent perturbations.

Another potential source of bias in the above results concerns the power of tests with polarized mutations (Fu and Li's *D* and Fay and Wu's *H*), which is likely to be overestimated. We assume here that the ancestral state is known without ambiguity. This is generally not the case in practice, where additional information from a close outgroup is commonly used to infer the ancestral state. On one hand, using too close an outgroup may lead to difficulties if there are shared ancestral polymorphisms. On the other hand, especially with more distant outgroups, there may be a problem with homoplasy. It is not usually possible for all sites to infer the ancestral state as there may be a third state for the outgroup (e.g., if the intraspecific

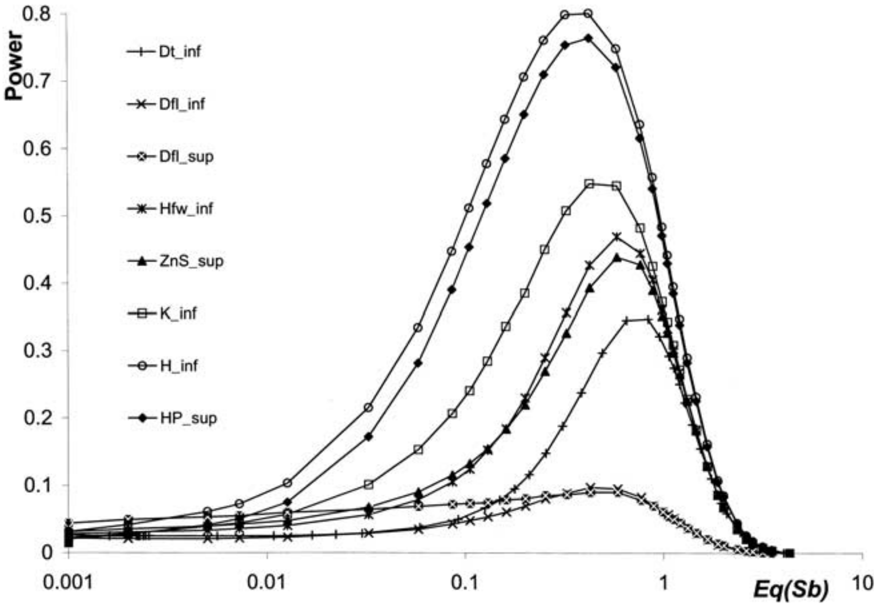

Fig. 6. Power of selective sweeps ($\alpha = 10{,}000$, $T_b = 0.001$, $\varepsilon = 0.001$; dashed curves) for a recent perturbation ($T_b = 0.001$) as a function of the strength of the perturbation computed according to Equation 3. Other characteristics are as in Fig. 4.

sample is polymorphic for A/G and the outgroup shows a C, a T, or a deletion on this site). Even if there are only two states, they may result from homoplasy (parallelism or reversion), potentially leading to nonrobust tests. The frequency of homoplasies is usually estimated assuming a constant mutation rate, with no mutational bias (Fay and Wu 2000) and it is thus likely to be underestimated. Another point about Fay and Wu's statistics is that, although designed to detect hitchhiking, it can show substantial power to detect bottlenecks (Fig. 4 versus Fig. 6; looking at the peaks of power, getting a significant *H* is only three times more likely under a hitchhiking than under a bottleneck). Thus, as for other tests, significant values of this test do not clearly distinguish between hitchhiking and bottleneck effects and do not provide strong evidence for selection. Relying exclusively on the frequency distribution of haplotypes to distinguish bottlenecks from hitchhiking effects may be unwise. This distribution could be affected by a number of other events. To make this distinction we prefer to rely on multilocus approaches in which selection is the only likely explanation (Hudson et al. 1987; Galtier et al. 2000). The expected effect of hitchhiking depends on the age of the last selective event in the vicinity of the marker considered, on its genetic distance *c* from the selected locus and on the selection coefficient *s*. The latter two effects can be summarized with little loss of information in terms of the ratio of *c* over *s* (Stephan et al. 1992; Depaulis et al., in press). Therefore, the expected effect of hitchhiking depends on the locus considered. On the contrary, the expected effect of a bottleneck of a given age and strength is the same whatever the locus considered. However, as seen above (Fig. 3), for some intermediate ranges of parameter values, the distribution of neutrality test statistics can be broadened, and they can show departure in both directions. In other words, the variance between loci can be drastically enhanced. A practical consequence is that, in multilocus studies, finding departure from neutrality tests in opposite directions for various loci does not exclude unambiguously bottleneck effects and does not provide strong evidence in favor of selection. It may thus be more appropriate to model explicitly both hitchhiking and bottlenecks, and to compare the likelihoods of the two models for a given set of multilocus data with, e.g., likelihood ratio tests (Galtier et al. 2000). However, such methods are not free from difficulties. The bottleneck alternative hypothesis is modeled as above, with a single age and strength parameter values whatever the locus considered. The alternative hitchhiking hypothesis is approximated with the same bottleneck approximation on each locus, but it allows the bottleneck parameter values to vary from one locus to the next. This makes the two models nested and allows for likelihood ratio tests to compare them. The difference we describe between bottleneck and hitchhiking for a given locus may have little effect on such multilocus analyses approximating selective sweeps by bottlenecks. Such a method probably uses primarily the information contained in the difference between loci: under hitchhiking effects, various loci support different ages and strengths of the perturbation. The method is computationally demanding and its properties have therefore been little investigated, but it seems to behave reasonably (Galtier et al. 2000). However, the robustness with respect to its assumptions is unknown. In particular, the asymptotic χ^2 approximation for testing the significance of the difference between the two models has not been thoroughly evaluated. Intragenic recombination has not yet been implemented in the method, and it is thus applied to regions of the datasets showing no evidence for recombination. However, most recombination events may not be detected (Hudson and Kaplan 1985) and intragenic recombination is known to affect maximum likelihood methods (Schierup and Hein 2000). But it is practically difficult to implement using full likelihood methods. Future work should focus on a more tractable likelihood approach based on a set of summary statistics that may capture most information (see, e.g., Wall 2000).

Acknowledgments. We thank J. Baines, M. Cobb, N. Galtier, and A. Navarro and anonymous reviewers for helpful discussions and comments on the manuscript. A user-friendly version of the haplotype tests is available at http://www.snv.jussieu.fr/mousset/. Our research is supported by CNRS UMR 7625, GDR 1928 and the Ecole Pratique des Hautes Etudes.

References

Andolfatto P, Wall JD, Kreitman M (1999) Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics 153:1297–1311

Barton NH (1998) The effect of hitch-hiking on neutral genealogies. Genet Res 72:123–133

Barton NH (2000) Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 355:1553–1562

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–796

Depaulis F, Mousset S, Veuille M (2001) Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol Biol Evol 18:1136–1138

Depaulis F, Mousset S, Veuille M (in press) Detecting selective sweeps with haplotype tests. In: Nurminsky D (ed) Selective sweep. Landes Bioscience, Georgetown, USA

Depaulis F, Veuille M (1998) Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol 15:1788–1790

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Pop Biol 3:87–112

Fay JC, Wu CI (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol Biol Evol 16:1003–1005

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Felsenstein J (1992) Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet Res 59:139–147

Fu YX (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143:557–570

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Galtier N, Depaulis F, Barton NH (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155:981–987

Hudson RR (1993) The how and why of generating gene genealogies. In: Takahata N, Clarck AG (eds) Mechanism of molecular evolution. Japan Scientific Societies Press, Sinauer Associates, Inc., Sunderland, MA, pp 23–36

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. Genetics 136:1329–1340

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Kelly JK (1997) A test of neutrality based on interlocus associations. Genetics 146:1197–1206

Nordborg M, Tavaré S (2002) Linkage disequilibrium: What history has to tell us. Trends Genet 18:83–90

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189

Schierup MH, Hein J (2000) Recombination and the molecular clock. Mol Biol Evol 17:1578–1789

Stephan W, Wiehe THE, Lenz MW (1992) The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor Popul Biol 41:237–254

Strobeck C (1987) Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. Genetics 117:149–154

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

Wall JD (1999) Recombination and the power of statistical tests of neutrality. Genet Res 74:65–69

Wall JD (2000) A comparison of estimators of the population recombination rate. Mol Biol Evol 17:156–163

Wall JD, Hudson RR (2001) Coalescent simulations and statistical tests of neutrality. Mol Biol Evol 18:1134–1135

Watterson GA (1975) On the number of segregation sites. Theoret Popul Biol 7:256–276