Measuring the degree of starshape in genealogies - summary statistics and demographic inference

Genet. Res., Camb. (2009), **91**, *pp.* 281–292. © Cambridge University Press 2009
doi:10.1017/S0016672309990139 Printed in the United Kingdom

Measuring the degree of starshape in genealogies – summary statistics and demographic inference

KONRAD LOHSE* AND JEROME KELLEHER
Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK

(*Received 27 August 2008 and in revised form 4 February 2009*)

Summary

The degree of starshape of a genealogy is readily detectable using summary statistics and can be taken as a surrogate for the effect of past demography and other non-neutral forces. Summary statistics such as Tajima's *D* and related measures are commonly used for this. However, it is well known that because of their neglect of the genealogy underlying a sample such neutrality tests are far from ideal. Here, we investigate the properties of two types of summary statistics that are derived by considering the genealogy: (i) genealogical ratios based on the number of mutations on the rootward branches, which can be inferred from sequence data using a simple algorithm and (ii) summary statistics that use properties of a perfectly star-shaped genealogy. The power of these measures to detect a history of exponential growth is compared with that of standard summary statistics and a likelihood method for the single and multi-locus case. Statistics that depend on pairwise measures such as Tajima's *D* have comparatively low power, being sensitive to the random topology of the underlying genealogy. When analysing multi-locus data, we find that the genealogical measures are most powerful. Provided reliable outgroup information is available they may constitute a useful alternative to full likelihood estimation and standard tests of neutrality.

1. Introduction

The motivation for studying the impact of past demography on sequence data is two-fold. Firstly, changes in population size are interesting in their own right, being intimately linked to processes such as speciation or geographic range shifts. Secondly, the standard neutral model (SNM) of a randomly mating Wright–Fisher population of constant size and discrete generations, hardly ever describes the patterns of diversity found in natural populations. Thus, studies aiming to detect loci under selection are faced with the considerable challenge of fitting realistic demographic models against which selection can be tested e.g. Glinka *et al.* (2003), Hamblin *et al.* (2004), Haddrill *et al.* (2005), Ometto *et al.* (2005) and Thornton & Andolfatto (2006). Since the rate of coalescence is inversely proportional to the effective population size, it is clear that demographic changes must leave a detectable signature in genealogies (Felsenstein, 1992). In general, positive population growth distorts genealogies towards a starshape with shorter internal branches, resulting in more low frequency variants and a unimodal rather than multi-peaked mismatch distribution (Slatkin & Hudson, 1991; Harpending, 1994; Schneider & Excoffier, 1999). In contrast to selective processes that act on single genetic variants, demography affects the whole genome, so one expects to find a concordant signature across loci (Tajima, 1989; Galtier *et al.*, 2000).

Approaches to demographic inference fall into three broad categories; for a review see Emerson *et al.* (2001). Firstly, likelihood methods, which are available for bottleneck and exponential growth models, make use of all the information in a sample by integrating over a large set of likely genealogies (Griffiths & Tavaré, 1994; Kuhner *et al.*, 1995). Although optimal in terms of statistical power and accuracy, likelihood estimation is computationally intensive and requires a fully specified alternative model. Therefore realistic growth histories often remain analytically intractable. Secondly, there are tree-based methods, which take the branch length information of a reconstructed tree as their starting point. Assuming that

* Corresponding author. Tel: +44 (0)131 650 5508. e-mail: K.R.Lohse@sms.ed.ac.uk

sequence evolution is clock-like, the number of lineages can be plotted against time and the shape of this trajectory compared with its neutral expectation (Nee *et al.*, 1995; Pybus *et al.*, 2002). Despite their conceptual appeal, these methods neglect any uncertainty in tree topology and are thus only as good as the reconstructed tree they are based on. Furthermore they cannot deal with recombination by definition. Finally, there are classical neutrality tests, most of which do not explicitly consider the genealogy but instead use more immediate aspects of the data such as the frequency spectrum of mutations, e.g. Tajima's D (Tajima, 1989) and Fu & Li's D (hereafter referred to as D_2) (Fu & Li, 1993), the haplotype distribution, e.g. Fu's F_S (Fu, 1996; Innan *et al.*, 2005), or the mismatch distribution, e.g. the raggedness statistic (Slatkin & Hudson, 1991). Compared with likelihood estimation, summary statistics are straightforward to calculate and their distribution can be simulated under almost any growth model.

Considering the zoo of statistics available and their wide use, there are surprisingly few studies that systematically compare their power, and those that do mainly consider bottlenecks and single locus data (Simonsen *et al.*, 1995; Fu, 1996; Ramos-Onsins & Rozas, 2002; Depaulis *et al.*, 2003; Ramirez-Soriano *et al.*, 2008). However, joint analysis of multiple loci is not only necessary to distinguish between selective and demographic events (Galtier *et al.*, 2000) but also potentially far more powerful than inferences based on a single locus. An added advantage of multi-locus analysis is that both means and variances of summary statistics can be used for testing. Variance based tests were first developed for microsatellite data (Di Rienzo *et al.*, 1998; Reich *et al.*, 1999) but are now routinely used to analyse sequence data from multiple loci (Pluzhnikov *et al.*, 2002; Haddrill *et al.*, 2005; Heuertz *et al.*, 2006) or even species (Hickerson *et al.*, 2006).

A general conclusion that has emerged from simulation studies is that tests based on the number and distribution of haplotypes have more power to detect bottlenecks than statistics based on π, in particular Tajima's D (Ramos-Onsins & Rozas, 2002; Innan *et al.*, 2005; Ramirez-Soriano *et al.*, 2008). Earlier, Felsenstein made a theoretical argument for the inferiority of pairwise measures (Felsenstein, 1992). Their large variance under neutrality arises both from their sensitivity to the last coalescence event and the random genealogical topology (Tajima, 1983). Under the SNM more symmetric genealogies are on average associated with higher π and more ragged mismatch distributions than asymmetric genealogies. It is important to realize that this topological variance is independent of the already large variance in coalescence times inherent in the genealogical process. In other words 'despite their aura of robustness' (Felsenstein, 1992), statistics based on π suffer from an unnecessarily large variance under neutrality, and hence have comparatively low power. Despite these results, D and mismatch distributions continue to be the methods of choice for demographic inferences in population genetics and phylogeography, respectively.

Following Felsenstein's recommendation that 'there is much to gain from explicitly taking the genealogical relationship of a sample into account' (Felsenstein, 1992), the aim of this study is to consider how genealogical information can be used for demographic inference in a summary statistics framework. Our premise here is that the mutation rate is sufficiently high relative to the per site recombination rate such that non-recombining blocks of sequences can be easily identified and treated as independent loci.

Given that there is usually not enough information in within-species sequences data to infer the full topology unambiguously it seems important to ask which part of the topology yields most information. The first part of the paper introduces some simple measures of starshape, which are based on the properties of a rooted genealogy. Using simulations, their power to detect a history of exponential growth is compared with standard neutrality tests for both the single and multi-locus cases. We focus on the exponential growth model for two reasons. Firstly, although it is a frequently used demographic model, the power of summary statistics to detect exponential growth has been little investigated. Secondly, likelihood methods are available, which can be taken as an absolute 'upper bound' of power for comparison. Such a direct comparison between summary statistics and the optimal likelihood methods is lacking so far.

2. Summary statistics

Several neutrality tests compare two different estimators of the scaled mutation rate (Fu & Li, 1993; Tajima, 1989; Fay & Wu, 2000) $\theta = 4N_e\mu$, where μ is the mutation rate and N_e the effective population size, which capture different aspects of the data. Most prominently, Tajima's D is defined as the difference between θ estimated as π, and $\theta_w = S/a_n$ (Watterson's θ, where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$, n is the sample size and S the total number of polymorphic sites in the sample), normalized by the standard deviation of this difference. Genealogies from growing populations typically have relatively more low frequency variants and hence tend to have a negative D.

While neutrality tests are commonly based on the frequency spectrum and π, it is instructive to consider departures from the SNM in terms of their effect on the genealogy. Such tree-thinking necessarily underlies summaries that make use of outgroup information, e.g. D_2 has a straightforward genealogical interpretation. Below two different ways of employing

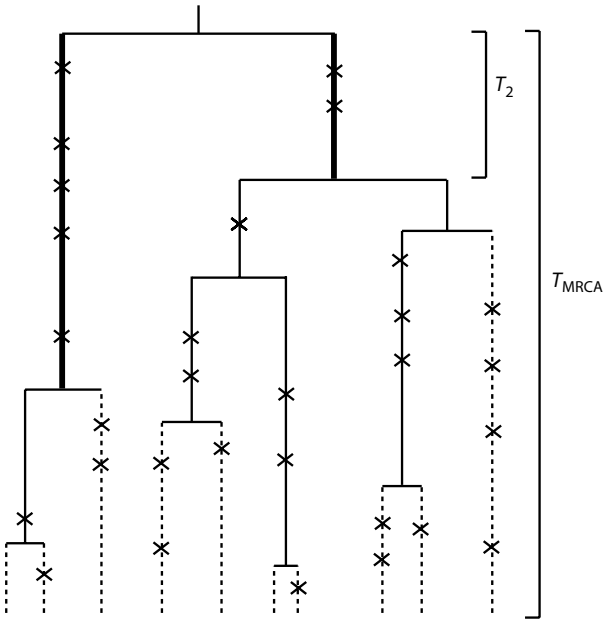

Fig. 1. Random genealogy of a sample of 20 sequences. The root partitions the sample into two subclades of size 3 and 7. Rootward branches are shown as bold, terminal branches as dotted lines, mutations are represented as crosses. The time interval until the last coalescence event, T_2, is shorter than average under the SNM. In this example $S=30$, $\eta_R=7$, $\eta_{R\min}=2$ and $\eta_e=14$.

genealogical information in the construction of summary statistics are considered.

(i) *Genealogical ratios*

The rationale behind D_2 is to distinguish between two classes of mutations: those found on terminal branches, η_e and those on internal branches, η_i (Fig. 1) (Fu & Li, 1993). Suppose that some limited topological information can be inferred from the data. In particular, we will for now assume that the placement of the root is known. It is then possible to distinguish mutations found on the two rootward branches, which we shall denote η_R. Under the infinite sites assumption, these are all derived mutations that are shared by all individuals in either of the two sub-clades defined by the root. The advantage of considering the proximity of mutations to the root rather than the tips is twofold: firstly, rootward branches cover a greater proportion of the time to the most recent common ancestor of the sample (T_{MRCA}) and should, in general, be more informative about past changes in population size. Under the SNM, on average half of the T_{MRCA} is taken up by the coalescence of the last two lineages (T_2) (Fig. 1), whereas in a growing population, the smaller population size in the past forces the last two lineages to coalescence much more rapidly. Secondly, the average length of a branch connected to the root is less dependent on the sample size than the average length of a terminal branch.

Ideally, one wants to know the total number of mutations that have occurred during T_2, rather than the number of mutations on both rootward branches, η_R which is larger and depends on the topology, i.e. the order of the first node on the longer of the two branches (Uyenoyama, 1997, Appendix).

One possibility is to only consider the shorter of the two rootward branches that has exactly length T_2. Thus the number of mutations found on this branch, $\eta_{R\min}$, over θw constitutes a very simple measure of starshape.

$$X=\frac{\eta_{R\min}}{\theta_w}. \tag{1}$$

Such genealogical ratios have first been employed to study the effect of balancing selection on plant incompatibility loci (Uyenoyama, 1997). Being based on a single random event, X clearly neglects much of the information contained in the genealogy. Its power is limited by the probability of observing $\eta_{R\min}=0$ under neutrality. In other words, X is unlikely to be of much use in the case of a single locus.

Alternatively, one can ignore the uncertainty in node order and take the number of mutations found on both rootward branches relative to θ_w:

$$X_1=\frac{\eta_R}{\theta_w}. \tag{2}$$

It is possible of course to construct various composite measures from the number of mutations found on different parts of the genealogy. Here, we only consider one additional statistic, the relative difference between rootward and terminal mutations:

$$X_2=\frac{\eta_R-\eta_e}{\theta_w}. \tag{3}$$

The X statistics assume some knowledge of the tree topology that is usually unknown. Of course one could use some standard method of tree reconstruction and infer η_R and $\eta_{R\min}$ from the most likely topology. However, not only is it inefficient to reconstruct the full topology when all that is required is the placement of the root, conditioning on a single tree also ignores any topological uncertainty. We have therefore developed a simple scheme of inferring the root in a sample of polarized sequences that circumvents these problems.

Under the infinite sites assumption, a necessary criterion for the root-node is that no mutations are shared between the two subsets on either side. One can show that if both branches connected to the root carry mutations, i.e. $\eta_{R\min}>0$ there exists exactly one bipartition of the sample with no mutational overlap. If however one or both of the rootward branches of

the genealogy carry no mutations there may be multiple bipartitions that meet this criterion. In this case $\eta_{R\min}=0$ and the tree reconstructed from such a sample would have an unresolved polytomy at its base. To incorporate the topological uncertainty about the placement of the root, we compute the average value of η_R over all partitions that are compatible with the criterion of no mutational overlap. Note that in contrast to most tree reconstruction algorithms that join similar sequences (i.e. start from the tips down the tree), our scheme is divisive (i.e. it starts from the root). To avoid having to consider all possible bipartitions of the sample ($2^{n-1}-1$), we make use of the fact that any sequences that share mutations have to be on the same side of the root. By first binning sequences that share at least one mutation, we can directly calculate η_R and the number of possible partitions.

(ii) *Starting from the limiting case*

A different approach is to construct summaries that measure departures from the limiting case of a perfectly star-shaped genealogy. Star-shaped genealogies have some convenient properties that can be used for this. Assuming that outgroup information is available, one can record the number of terminal mutations in each sequence i (because lineages are exchangeable, the labelling is arbitrary), V_i. In a perfectly star-shaped genealogy, all mutations must fall onto terminal branches by definition. Thus one expects the number of derived mutations in a sequence to be half the average pairwise diversity, i.e. $E[V_i]=\pi/2$. The statistic R_{2E} proposed by Ramos-Onsins and Rozas measures the average departure from this expectation:

$$R_{2E}=\frac{\left(\sum_{i=1}^{n}\left(V_i-\frac{\pi}{2}\right)^2/n\right)^{1/2}}{S} \tag{4}$$

(Ramos-Onsins & Rozas, 2002, eqn (2)). R_{2E} has proven superior to a wide range of summary statistics in detecting histories of bottlenecks (Ramos-Onsins & Rozas, 2002). However, because of its dependence on π, one may suspect it to suffer from a large variance under neutrality. We therefore consider a similar statistic that uses the observed S rather than π to assess the degree of starshape. Consider the total number of derived mutations in each sequence, D_i. Note that $\sum_{i=1}^{n} D_i=\sum_{i=1}^{n-1} i\xi_i$, in terms of the unfolded frequency spectrum, where ξ_i denotes derived mutations that occur i times in the sample. Using the fact that $E[D_i]=S/n$ in a star-shaped genealogy we can define a new statistic:

$$R_S=\frac{\left(\sum_{i=1}^{n}\left(D_i-\frac{S}{n}\right)^2/n\right)^{1/2}}{S}. \tag{5}$$

Since under neutrality a large proportion of mutations will be found on inner branches, i.e. be shared by many sequences, $E[D_i]=S/n$. In other words, R_S is such that smaller values are expected under a history of growth.

3. Methods

(i) *Summary statistics and demographic model*

We carried out coalescent simulations in ms (Hudson, 2002) to compare the power of a range of summary statistics to distinguish between the SNM and a history of exponential growth. In addition to D, D_2, R_{2E} and the new statistics defined above, F_S, (Fu, 1996) and H (Fay & Wu, 2000) were considered. F_S is based on the number of haplotypes in the sample and has previously been found to be more powerful than statistics based on the frequency distribution (Fu, 1996; Ramos-Onsins & Rozas, 2002). H was conceived as a test for the effect of selection on linked neutral sites (Fay & Wu, 2000) and is not expected to have power to detect continuous growth. However, other demographic scenarios such as moderate bottlenecks may perturb genealogies in ways similar to genetic hitchhiking resulting in significant values of H. We assume that the population size has grown exponentially with rate α to its current size N_0:

$$N(t)=N_0\ \mathrm{e}^{-\alpha t}. \tag{6}$$

Following standard practice, this exponential growth is incorporated through a re-scaling of time (Slatkin & Hudson, 1991). We define a rescaled time T_{coal} relative to N_0 and α:

$$T_{\mathrm{coal}}=\int_0^t \frac{\mathrm{e}^{\alpha t}}{2N_0}\mathrm{d}t=\frac{(\mathrm{e}^{\alpha t}-1)}{2N_0\alpha}. \tag{7}$$

This represents the total amount of genetic drift that has occurred. It is convenient to define a growth rate relative to N_0 as $A=2N_0\alpha$, which gives:

$$T_{\mathrm{coal}}=\frac{\mathrm{e}^{A t/2N_0}-1}{A}. \tag{8}$$

(ii) *Power test*

Critical values of 5% confidence for each statistic were determined from 10 000 replicate genealogies simulated under the SNM for each of a wide range of S values (1–250) (Hudson, 1993; Braverman *et al.*, 1995; Ramos-Onsins *et al.*, 2007). Genealogies from growing populations were simulated conditional on θ. For each replicate the alternative hypothesis of positive growth was tested by comparing the observed value of a statistic to the critical value given the observed S. Power was estimated as the proportion of

10 000 replicate genealogies for which a statistic was below its critical value in a one-tailed test. Power to reject the SNM was recorded for a large range of parameter combinations. We compared the performance of statistics for different growth rates, ($0 < A < 50$), sample sizes ($n = 10$, 50) and values of θ (5–50). When varying θ, we chose a fixed value of $A = 8$. This seems compatible with growth rates estimated from empirical data. For example, variation at silent sites in the *Adhr* region and X-linked genes in *Drosophila pseudoobscura* is consistent with $A = 7$ (Schaeffer, 2002). While θ can be arbitrarily high for mitochondrial data, $\theta = 20$ may be unrealistic for nuclear loci in out-crossing species. Therefore, power was evaluated for a range of θ values (5–50) again keeping the growth rate fixed at $A = 8$.

When using means and variances of summary statistics across loci, power was determined analogously to the single locus case. Critical values of 5% confidence of means and variances of statistics were determined from 10 000 replicate sets of loci with the exact same combination of S values. Although computationally expensive, this avoids making any assumptions about the distribution of mutation rates between loci. However, given that mutation rates vary along the genome assuming the same θ for all loci to simulate the alternative history of growth seems unrealistic and may lead to overestimation of power. We checked for the influence of heterogeneity in mutation rates on power by repeating the multilocus power tests with θ values drawn from a gamma distribution with $\alpha = 2$ (Pluzhnikov *et al.*, 2002) and a scale parameter equivalent to a mean of $\theta = 20$. This combination of growth and mutation rates is roughly comparable to mutation rate estimates for nuclear loci in *Drosophila melanogaster* (Galtier *et al.*, 2000). As before we assumed no recombination within loci as well as absence of linkage between loci, i.e. replicate genealogies were simply treated as multiple loci.

(iii) *Likelihood method*

In practice, both θ and A are unknown, and their likelihood should, in principle, be estimated jointly. However, because of the non-independence of these two parameters, this is not a practical option. Following standard practice we alternated between maximum likelihood estimation of A and θ (Griffiths & Tavaré, 1994). First a maximum likelihood estimate (MLE) for θ under the SNM was estimated using the program GENETREE (http://www.stats.ox.ac.uk/griff/software.htm). In a second step, this MLE for θ was fixed to run a likelihood surface for A. Finally, the MLE value for A was used to re-evaluate θ. This scheme yields two MLEs for θ for each replicate, one under the assumption of no growth and one given the most likely growth rate, which were compared in a likelihood ratio test (LRT). We did not find that the MLE estimates for A and θ improved upon repeated re-evaluation suggesting that a single round of estimation is sufficient for this moderate growth scenario. 100,000 runs were performed for each likelihood surface evaluation. Again, the proportion of replicate genealogies for which the null hypothesis could be rejected was taken as a measure of statistical power. Due to the long computing time, 100 replicates per parameter combination were used.

4. Results

(i) *Single locus*

In general, both the likelihood method and summary statistics have low power to detect a history of moderate ($A < 8$) exponential growth for $n = 10$ (Fig. 2). As expected, the likelihood method is most powerful overall, although its superiority is surprisingly small. For example, based on the LRT the SNM is rejected for 30% of genealogies simulated under exponential growth of $A = 4$. In comparison, R_S and R_{2E} detect this history of growth in 23% of cases (Fig. 2).

Consistent with previous results, F_S, R_{2E}, and the new measure R_S, are considerably more powerful than both D and D_2 (Ramos-Onsins & Rozas, 2002; Ramirez-Soriano *et al.*, 2008). For $\theta = 20$, F_S is the most powerful statistic. The new measure R_S has consistently higher power than R_{2E}. As expected, H and X have no power to distinguish between the SNM and the growth case (not shown). However, the other two genealogical ratios perform surprisingly well. X_1 has higher power than D_2 and the power of X_2 is between that of R_{2E} and R_S (Fig. 2). The complete lack of power of D for $n = 10$ is somewhat surprising. Comparison with the result for $n = 50$ (Fig. 3) reveals that its performance is strongly dependent on sample size. We ran additional simulations (not shown) and found that for $n < 15$ extremely negative values of D are more likely under neutrality than under growth resulting in a rejection rate of the SNM of less than 5%. In other words, when n is small, the variance of D under neutrality is too large to detect exponential growth.

In general, all statistics have considerably higher power for $n = 50$ (Fig. 3). Interestingly, it never reaches 100% even when growth is extreme ($A = 50$). However, the relative effect of the sample size on power differs between statistics. For instance, X_1 improves relatively little in comparison to other measures. This is to be expected given that even small samples are likely to include the deepest split in the genealogy of the whole population (Saunders *et al.*, 1984). For $n = 10$, the power of all statistics decreases

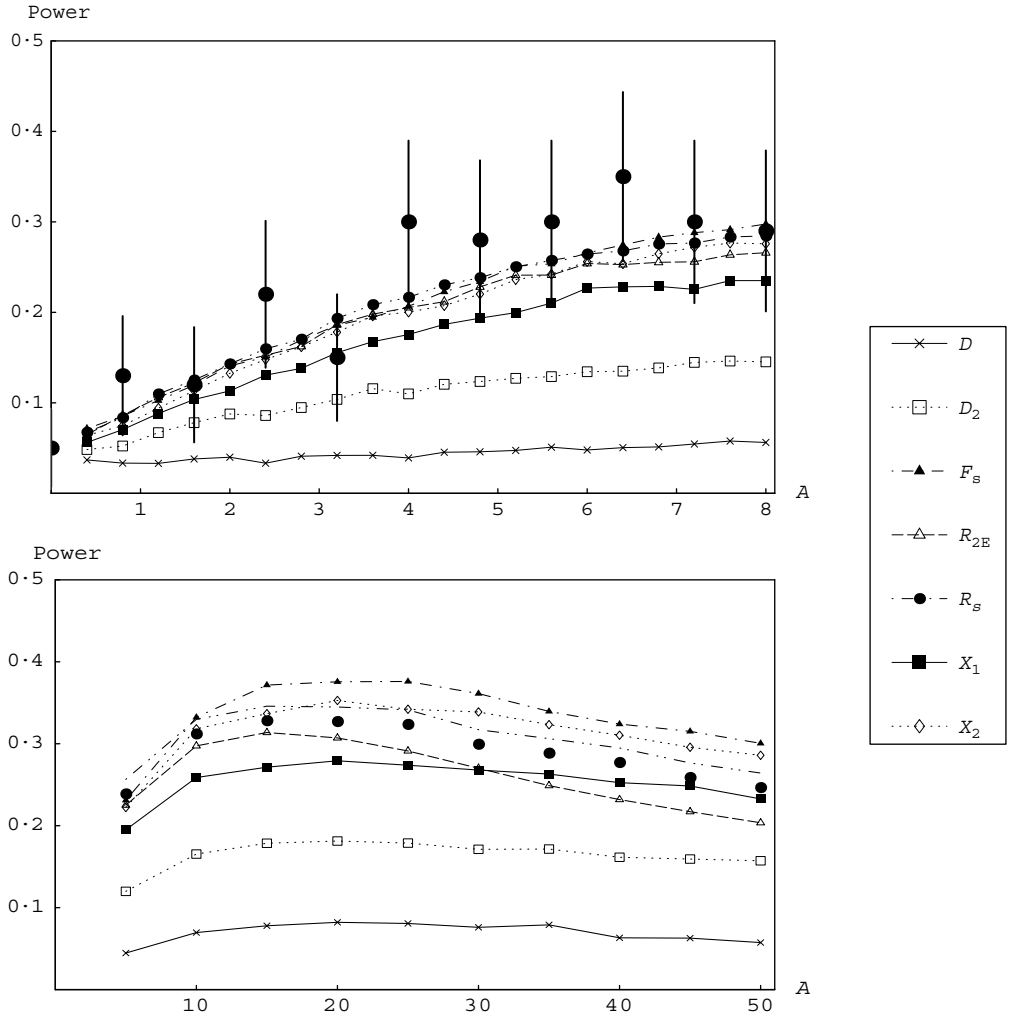

Fig. 2. Power of summary statistics and likelihood method against exponential growth rate $A = 0$–50. $n = 10$, $\theta = 20$. Each point is based on 10 000 replicate simulations. The power of the likelihood method was estimated from 100 replicates (see large filled circles and error bars).

for histories of extreme growth ($A > 25$) (Fig. 2). This is due to the overall shortening of genealogies under rapid growth.

The mutation rate has a relatively small influence on power. In general, the power of all measures increases with θ (Fig. 4). However, the trajectories X_1 and F_S level off while the power of the other statistics continues to improve with increasing values of θ. The power of F_S is limited by the number of haplotypes (which cannot exceed n).

To check how statistics are affected by the topological variance, genealogies simulated under the alternative history of growth were sorted according to the bipartition by the root and the proportion of significant values determined for each topology class. Figure 5 clearly shows that the two statistics based on π, D and R_{2E} as well as D_2 are sensitive to asymmetric topologies. The chance of observing a significant value increases markedly with topological asymmetry. This effect is most pronounced for D, which has no 'power' to reject the SNM unless genealogies are very asymmetric and growth is weak. In contrast, the dependency of X_1 on the rootward partition is relatively slight and in the opposite direction, i.e. the chance of rejecting the SNM is smaller for asymmetric genealogies (Fig. 5).

(ii) *Multiple loci*

Compared with the relatively subtle effect both θ and n have on statistical power, increasing the number of loci improves power dramatically. In the mean-based test, all statistics apart from D have a power of close to 100% to detect a history of moderate exponential growth ($A = 8$) for 10 loci. However, the relative performance of statistics changes slightly compared with the single locus case. Notably, X_2 has higher power than all other summary statistics (Fig. 6). The power of X is slightly lower than that of X_1 (not shown). Analogously to the results for a single locus, power

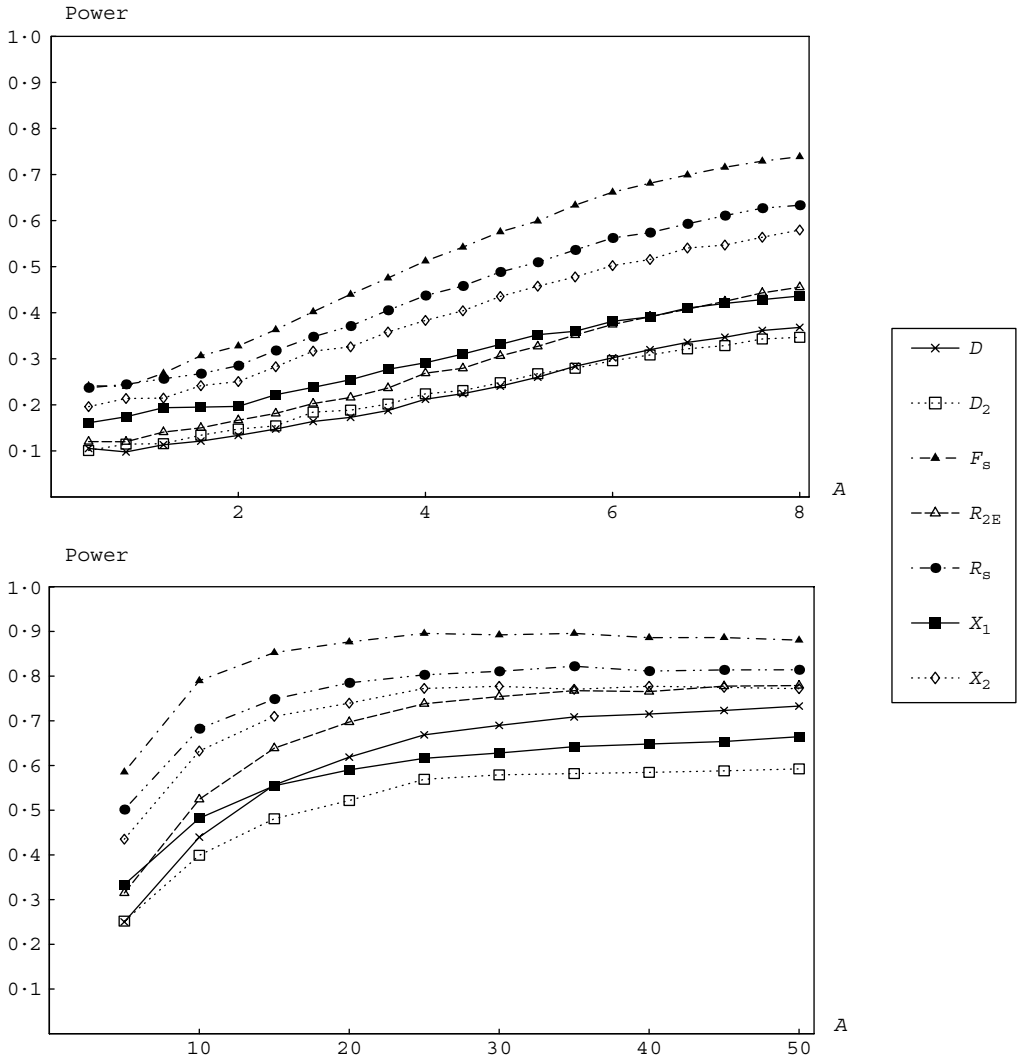

Fig. 3. Power of summary statistics against exponential growth rate $A=0$–50. $n=50$, $\theta=20$. Note the different range (0–1) on the y-axis compared with Fig. 2.

increases both with more extreme growth scenarios and larger n (not shown).

As one may suspect, the increase in power with the number of loci is slower for the variance test. More importantly, the relative performance of statistics is very different. By far the most powerful statistic in the variance test is X_1 followed by D and X (Fig. 7). This indicates a general trade-off. Statistics with a high variance under the SNM have comparatively low power in the single-locus case and the mean test, but high power in the variance test and vice versa.

Allowing for heterogeneity in mutation rates between loci affects both the relative performance of summary statistics and their overall power. As one may expect, heterogeneity in θ generally results in a decrease in power. In the mean-based test, the three X statistics are most affected. However, in the variance test the performance of X_1 is little affected. This statistic even has slightly higher power when mutation rates vary between loci. This appears to be due to the non-normal distribution of X_1 under growth. Genealogies with more than one possible root-partition generally have a very low value of X_1, since we take an average over all possible partitions most of which will be associated with $X_1=0$.

5. Discussion

It is important to distinguish between the general limitations that genealogical and mutational stochasticity impose on demographic inference from genetic data and problems associated with particular methods. Two main conclusions emerge from comparing the performance of the new 'genealogical statistics' to classical neutrality tests and the LRT.

(i) *General limits to demographic inference*

The signatures that changes in population size leave in genealogies are typically subtle compared with the

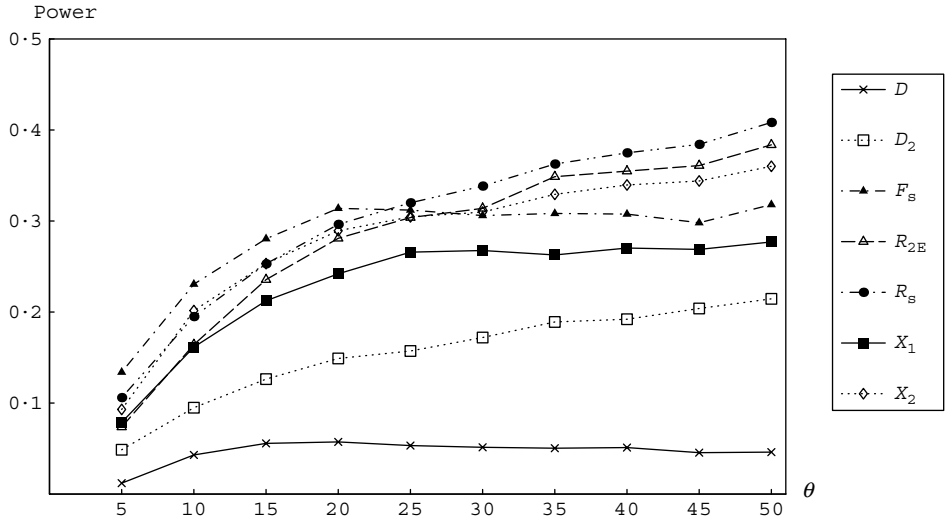

Fig. 4. Power of summary statistics to detect a history of exponential growth ($A=8$) against θ. $n=10$.

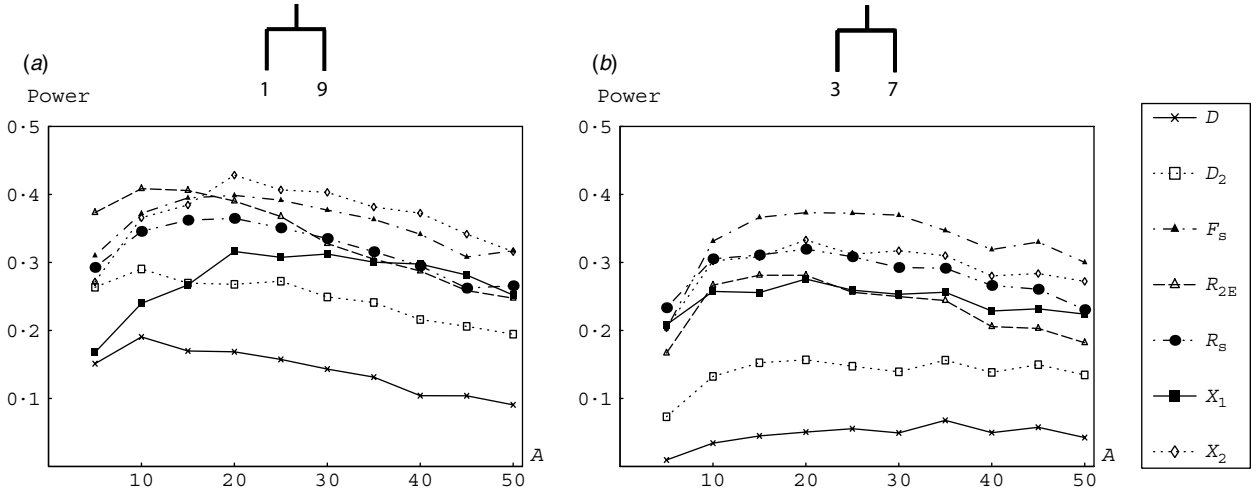

Fig. 5. The effect of topological asymmetry on statistical power (simulation parameters as in 2). Genealogies of Fig. 2 were sorted according to the partition by the root (shown above plot). Only the most asymmetrical partition (9, 1) (a) and one other case (7, 3) (b) are shown. Results for the other three partitions were very similar to (b). Note that since lineages are exchangeable all asymmetrical partitions have the same probability $P_a=2/(n-1)$ (Tajima, 1983, eqn (2)).

randomness of the ancestral process. Thus all methods have low power to distinguish between the SNM and histories of moderate growth in the single locus case. A surprising finding of this study was that the full likelihood method only works marginally better than the most powerful summary statistics. Changes in N_e disproportionally affect the length of the basal branches of a genealogy. However, because these rootward branches also contribute most to the variance in total tree length, inferences based on a single locus will be weak at best. It is telling that the X statistics which only considers the last coalescence events in the history, outperform standard neutrality tests in the variance test when multiple realizations of this event, i.e. loci, are available. As has been argued before, most statistical power can be gained by increasing the number loci, which represent independent realizations of the ancestral process, rather than the sample size or the length of sequence (Felsenstein, 1992; Kliman *et al.*, 2000; Wakeley, 2004).

(ii) *Pairwise measures*

Independent of the general limits to demographic inference, pairwise measures such as D have particularly low power to infer demography. This has been found in previous simulation studies, which consider other demographic scenarios such as strong bottlenecks and rapid logistic growth (Fu, 1996; Ramos-Onsins & Rozas, 2002; Ramirez-Soriano *et al.*, 2008). The fundamental flaw of pairwise measures can be best understood in terms of the underlying genealogy.

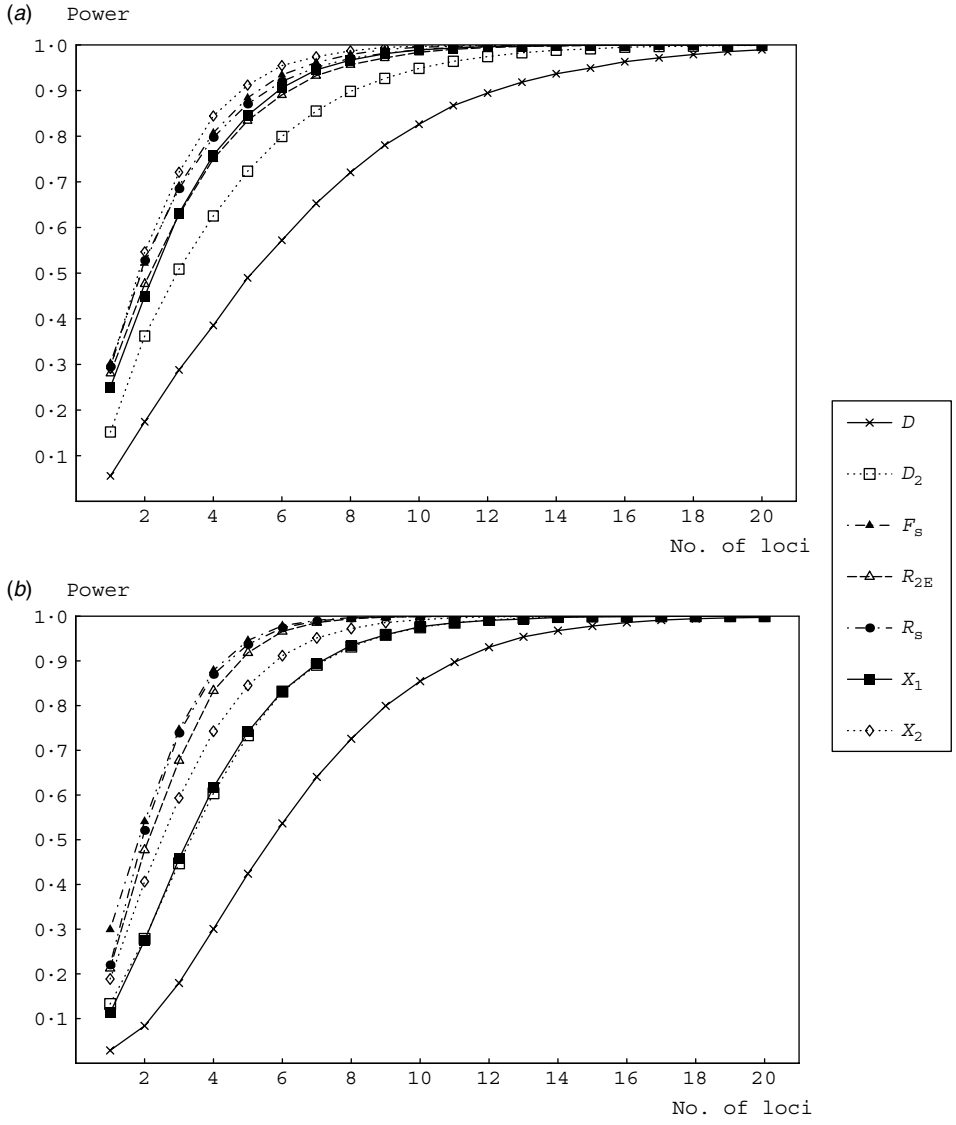

Fig. 6. Power of summary statistics to detect a history of growth $A=8$ using the mean across multiple loci against the number of loci, $n=10$ (A) and $\theta=20$ (B). Assuming mutational rate heterogeneity (θ gamma distributed with $\alpha=2$ and $E[\theta]=20$).

In contrast to selection and population structure, changes in N_e on their own only alter the distribution of branch lengths without affecting the topology, which can be regarded as a random nuisance parameter. While the full topology can rarely be reconstructed, there is potentially a lot of topological information in sequence data. Thus, the challenge that any efficient inference method has to meet is to separate this topological information from the relevant branch length information while taking topological uncertainty into account. Tree-based methods such as lineage-through time plots clearly fall short of the latter because they rely on a fully resolved topology. Pairwise measures on the other hand simply ignore the confounding effect of the topology (Felsenstein, 1992). It is thus easy to see why D has power only when sample sizes are large. While increasing sample size adds increasingly shorter external branches and therefore little additional information, it does reduce the chance of extremely asymmetric bipartitions by the root which are responsible for much of the variance in π and hence D.

Perhaps worryingly, this sensitivity to the topology not only translates into a loss of statistical power but also means that negative D values may in fact be more informative about the topological asymmetry of the genealogy (which may be caused by other non-neutral forces, e.g. selection) underlying the sample than about past growth. In order to distinguish between the effects of selection and demography, topology needs to be separated from branch length information. One approach is to explicitly account for the topology information if possible. For instance, one could determine confidence intervals of statistics conditional

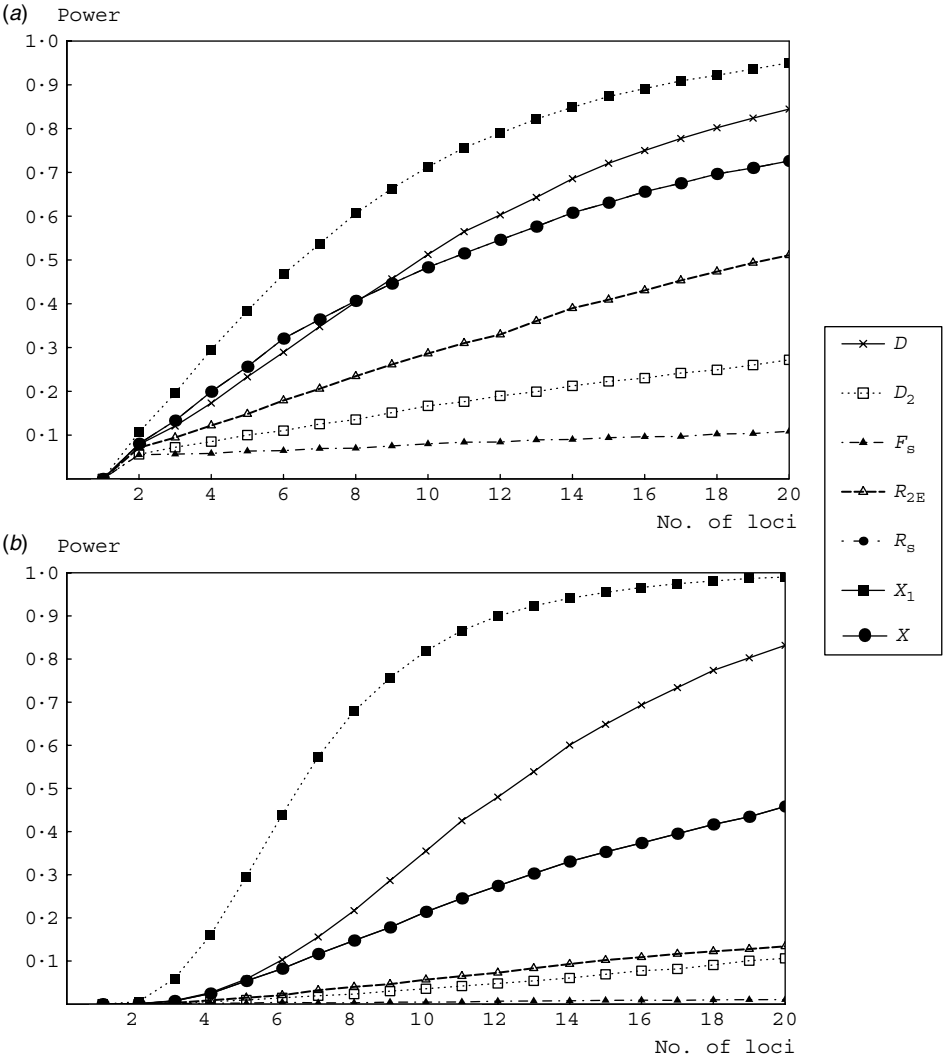

Fig. 7. Power of summary statistics in the variance-based tests across multiple loci for three different growth rates (from left to right $A=2$, 4, 8). (A) $\theta=20$. (B) Assuming mutational rate heterogeneity (θ gamma distributed with $\alpha=2$ and $E[\theta]=20$).

on the bipartition by the root if this is known. Not surprisingly, this improves the power of D, but has little effect on statistics that are not based on π (not shown). The alternative is to use measures that are less sensitive to the topology. F_S and other haplotype statistics have previously been shown to be more powerful than frequency spectrum statistics for this very reason (Depaulis *et al.*, 2003; Innan *et al.*, 2005). However, it has also been noted that F_S sometimes behaves erratically (Fu & Li, 1993; Ramos-Onsins & Rozas, 2002). As mentioned earlier, its power levels off with increasing θ (Fig. 4), because the sample size sets an upper bound to the number of haplotypes.

(iii) *Recombination and topological uncertainty*

The X statistics presented here fall somewhere in between tree-based methods and classical summary statistics. They exploit the fact that changes in population size disproportionally affect the relative length of the deepest branches in the genealogy and make use of topological information, without sacrificing the simplicity of the summary statistics framework. Given their high power in the multilocus case, how useful are such genealogical ratios in practice?

Recombination presents a fundamental problem to tree-based methods like the X statistics, which are defined only for non-recombining sequences. Similarly, likelihood methods that can deal with recombination are currently not available. To wrongly reconstruct trees from recombining data can potentially be severely misleading especially in the context of demographic inference. In fact, genealogical ratios similar to the ones presented here have been used to show that recombination can mimic the effect population growth has on the shape of inferred genealogies. Internal branches will appear relatively shorter and the tree overall more star-shaped (Schierup & Hein,

2000; Ramirez-Soriano *et al*., 2008). Ideally one would like to model recombination explicitly when making demographic inferences. However estimates of recombination rates are usually associated with a large uncertainty. Furthermore, it is notoriously difficult to distinguish between recombination and back-mutations.

One approach to circumvent these problems is to test for recombination beforehand (e.g. using the four gamete test) and exclude recombinant regions from the analysis if necessary. One can then both condition on there being no within-locus recombination and afford to use more powerful statistics such as the ones presented here. This strategy of identifying non-recombining stretches of sequence is increasingly used to analyse multilocus data, e.g. Galtier *et al*. (2000) or Jennings & Edwards (2005). Fortunately, many organisms appear to have lower recombination rates than model species such as Drosophila. For instance in a recent study on Australian birds only 6 out of 30 loci of intergenic sequence showed evidence for recombination (Jennings & Edwards, 2005). How profitable this scheme is ultimately depends on the relative magnitude and distribution of recombination and mutation rates. Before the genealogical ratios can be used on multiple loci, which have been pruned to exclude recombinant stretches, both the potential bias of such pruning and the effect of undetected recombination events on the genealogical ratios need to be properly evaluated. Interestingly, our method of inferring the root does in itself constitute a test for recombination and may help to focus on those recombination events that matter to the statistical test.

A related problem concerns the infinite sites assumption. Although the algorithm we have developed to compute the X statistics takes topological uncertainty into account, ignoring the possibility of back-mutations may underestimate the length of basal branches (Baudry & Depaulis, 2003). Although this source of error has been ignored here it should in principle be possible to account for back-mutations considering that they are independent of the assumptions of the genealogical process. In fact, any mutational model can be used to define statistics analogous to the genealogical ratios presented here. The problem with more complicated mutation models is in estimating the basal topology needed to calculate these measures.

(iv) *Conclusions*

In summary, the results confirm that only the most extreme demographic events leave a sufficient signature to be detectable in single locus data. Still, instead of the excessive and often non-quantitative employment of mismatch distributions, phylogeographic studies could benefit from using more powerful statistics such as R_S and R_{2E} to test demographic hypotheses. Conversely, population genetics studies of sequence data from multiple, unlinked loci could benefit from using summary statistics that incorporate genealogical information explicitly. When outgroup information is available and the assumptions of no within-locus recombination and infinite sites mutations can be justified, simple genealogical ratios are potentially more powerful than standard statistics. In taking the relative number of mutations found on specific parts of the genealogy as a measure of the degree of starshape, the demographic signal can be separated from irrelevant and confounding topological information. Extensions of this approach are feasible. For instance, one could consider the covariance between the number of basal and terminal mutations. Such simple statistics may be profitable for approximate likelihood or Bayesian approaches (Thornton & Andolfatto, 2006). There remains a need to understand the effect of pruning and undetected recombination events on tree reconstruction in general and tree-based measures such as the X statistics presented here in particular.

Many thanks to N. Barton, P. Haddrill, J. Polechova, D. Charlesworth, Kai Zeng and D. Obbard for helpful advice and discussion. Thanks also to Kelly Dyer for help with Genetree. Detailed comments and valuable suggestions from two anonymous reviewers on an earlier version of this manuscript greatly improved this work. K.L. is funded by a studentship from the Biotechnology and Biological Sciences Research Council. J.K. is funded by EPSRC.

References

Baudry, E. & Depaulis, F. (2003). Effect of misoriented sites on neutrality tests with outgroup. *Genetics* **165**, 1619–1622.

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.

Depaulis, F., Mousset, S. & Veuille, M. (2003). Power of neutrality tests to detect bottlenecks and hitchhiking. *Journal of Molecular Evolution* **57**, S190–S200.

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M. L., Haines, G. K. & Barch, D. H. (1998). Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics* **148**, 1269–1284.

Emerson, B. C., Paradis, E. & Thebaud, C. (2001). Revealing the demographic histories of species using DNA sequences. *Trends in Ecology and Evolution* **16**, 707–716.

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

Felsenstein, J. (1992). Estimating effective popuation size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59**, 139–147.

Fu, Y. X. (1996). New statistical tests of neutrality for DNA samples from a population. *Genetics* **143**, 557–570.

Fu, Y. X. & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

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

Glinka, S., Ometto, L., Mousset, S., Stephan, W. & De Lorenzo, D. (2003). Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* **165**, 1269–1278.

Griffiths, R. C. & Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions: Biological Sciences* **344**, 403–410.

Haddrill, P. R., Thornton, K. R., Charlesworth, B. & Andolfatto, P. (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research* **15**, 790–799.

Hamblin, M. T., Mitchell, S. E., White, G. M., Gallego, J., Kukatla, R., Wing, R. A., Paterson, A. H. & Kresovich, S. (2004). Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics* **167**, 471–483.

Harpending, H. C. (1994). Signature of ancient population growth in a low-resolution miitochondrial DNA mismatch distribution. *Human Biology* **66**, 591–600.

Heuertz, M., De Paoli, E., Kallman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M. & Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**, 2095–2105.

Hickerson, M. J., Dolman, G. & Moritz, C. (2006). Comparative phylogeographic summary statistics for testing simultaneous vicariance. *Molecular Ecology* **15**, 209–223.

Hudson, R. R. (1993). The how and why of generating gene genealogies. In *Mechanisms of Molecular Evolution* (Eds. N. Takahata & A. G. Clark), pp. 23–36. Sinauer, Sunderland, Mass.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

Innan, H., Zhang, K., Marjoram, P., Tavaré, S. & Rosenberg, N. A. (2005). Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics* **169**, 1763–1777.

Jennings, W. B. & Edwards, S. V. (2005). Speciation history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* **59**, 2033–2047.

Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., McCarter, J., Wakeley, J. & Hey, J. (2000). The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**, 1913–1931.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics* **140**, 1421–1430.

Nee, S., Holmes, E. C., Rambaut, A. & Harvey, P. H. (1995). Inferring population history from molecular phylogenies. *Philosophical Transactions of the Royal Society of London Series B* **349**, 25–31.

Ometto, L., Glinka, S., De Lorenzo, D. & Stephan, W. (2005). Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution* **22**, 2119–2130.

Pluzhnikov, A., Di Rienzo, A. & Hudson, R. R. (2002). Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* **161**, 1209–1218.

Pybus, O. G., Rambaut, A., Holmes, E. C. & Harvey, P. H. (2002). New inferences from tree shape: numbers of missing taxa and population growth rates. *Systematic Biology* **51**, 881–888.

Ramirez-Soriano, A., Ramos-Onsins, S. E., Rozas, J., Calafell, F. & Navarro, A. (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* **179**, 555–567.

Ramos-Onsins, S. E., Mousset, T., Mitchell-Olds, T. & Stephan, W. (2007). Population genetic inference using a fixed number of segregating sites: a reassessment. *Genetical Research* **89**, 231–244.

Ramos-Onsins, S. E. & Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* **19**, 2092–2100.

Reich, D., Feldman, M. & Goldstein, D. (1999). Statistical properties of two tests that use multilocus data sets to detect population expansions. *Molecular Biology and Evolution* **16**, 453–466.

Saunders, I. W., Tavaré, S. & Watterson, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–491.

Schaeffer, S. W. (2002). Molecular population genetics of sequence length diversity in the adh region of *Drosophila melanogaster*. *Genetical Research* **80**, 163–175.

Schierup, M. H. & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891.

Schneider, S. & Excoffier, L. (1999). Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089.

Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slatkin, M. & Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562.

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

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Thornton, K. & Andolfatto, P. (2006). Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**, 1607–1619.

Uyenoyama, M. K. (1997). Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics* **147**, 1389–1400.

Wakeley, J. (2004). Recent trends in population genetics: more data! more math! simple models? *Journal of Heredity* **95**, 397–405.