BMC Genetics

BioMed Central

Methodology article

Open Access

Fully Bayesian tests of neutrality using genealogical summary statistics

Alexei J Drummond*[1,2] and Marc A Suchard[3,4]

Address: [1]Bioinformatics Institute, University of Auckland, Private Bag 92019, Auckland, New Zealand, [2]Department of Computer Science, University of Auckland, Private Bag 92019, Auckland, New Zealand, [3]Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California, USA and [4]Department of Biostatistics, UCLA School of Public Health, Los Angeles, California, USA

Email: Alexei J Drummond* - alexei@cs.auckland.ac.nz; Marc A Suchard - msuchard@ucla.edu

* Corresponding author

Published: 31 October 2008

BMC Genetics 2008, **9**:68 doi:10.1186/1471-2156-9-68

Received: 8 February 2008
Accepted: 31 October 2008

This article is available from: http://www.biomedcentral.com/1471-2156/9/68

© 2008 Drummond and Suchard; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Many data summary statistics have been developed to detect departures from neutral expectations of evolutionary models. However questions about the neutrality of the evolution of genetic loci within natural populations remain difficult to assess. One critical cause of this difficulty is that most methods for testing neutrality make simplifying assumptions simultaneously about the mutational model and the population size model. Consequentially, rejecting the null hypothesis of neutrality under these methods could result from violations of either or both assumptions, making interpretation troublesome.

Results: Here we harness posterior predictive simulation to exploit summary statistics of both the data and model parameters to test the goodness-of-fit of standard models of evolution. We apply the method to test the selective neutrality of molecular evolution in non-recombining gene genealogies and we demonstrate the utility of our method on four real data sets, identifying significant departures of neutrality in human influenza A virus, even after controlling for variation in population size.

Conclusion: Importantly, by employing a full model-based Bayesian analysis, our method separates the effects of demography from the effects of selection. The method also allows multiple summary statistics to be used in concert, thus potentially increasing sensitivity. Furthermore, our method remains useful in situations where analytical expectations and variances of summary statistics are not available. This aspect has great potential for the analysis of temporally spaced data, an expanding area previously ignored for limited availability of theory and methods.

Background

The field of population genetics has a long history in the development of tests of selective neutrality. This is both because of the difficulty of developing a tractable alternative to the neutral theory and because of the ongoing debate about how well the neutral theory can explain real data. Although a number of important steps have been made to develop powerful tests of neutrality [1-3] there are evident problems with many currently available tests. For example many of the tests, such as Tajima's D (D_T) and Fu and Li's D (D_F) have difficulty in accurately discriminating between selection and changes in population size.

In fact, most available tests of neutrality can only test constant population size neutrality against alternatives that include both population growth and selection. Furthermore, most tests require accurate knowledge of the number of mutations that have occurred or the branch lengths in the gene tree, and do not adequately take into account the uncertainty in these quantities (i.e. most tests implicitly assume an infinite-sites model of evolution). Finally, tree-based summary statistics are often based on one estimate of the genealogy, despite the fact that the true genealogy and branch lengths are seldom known.

Broadly speaking, on the basis of the sequence information used, statistics for testing neutrality can be placed into three classes:

1. statistics that use the mutation (segregating site) frequency spectrum [1,2,4,5],

2. statistics that use the haplotype distribution [3,6,7] and

3. statistics that use the pair-wise distance (mismatch) distribution [8,9].

A recent comprehensive survey of the power of these different classes of tests for detecting population expansion found that classes 1 and 2 were generally more powerful than the best class 3 statistics [10]. Some of the best-known test statistics come from class 1 and essentially work by comparing aspects of the mutation frequency spectrum with neutral expectations. This class of test statistics include D_T [1], D_F [2] and the H statistic [5]. In the simplest case, these statistics can be used to measure deviations from the null hypothesis of constant population size, random mating and no recombination. For example D_F measures the normalized difference between the number of mutations on the external branches and the total number of mutations in the genealogy. Under the null hypothesis of neutral evolution the expectation of D_F is zero, and a significant departure from zero signifies selection (balancing, directional, negative), recombination or changes in population size. The last of these alternatives is problematic because exponential growth is expected to give results similar to directional or purifying selection. For this reason it would seem desirable to develop a method that directly accounts for alternative demographic models of population size through time. In this context, several studies have combined the use of summary statistics and demographic models [11-15].

Apart from biasing the mutation frequency distribution, selection may also affect the shape of the gene tree [16]. Although few attempts have been made to use this expectation in a rigorous test of neutrality (c.f. [17]), a number of branching models and summary statistics measuring tree imbalance exist in the literature of speciation models [18-21]. A method that could use information both from the mutation frequency spectrum and from the shape of the gene tree may be more powerful than either used individually.

If all sequences comprising a gene tree are sampled from the same time point (as is required by most tests of neutrality) then there is very little power to distinguish between selection and exponential growth. However if sequence data is available from different times, during which measurable evolution has taken place, as in RNA viruses and ancient mitochondrial DNA (mtDNA) data [22,23] then the power to distinguish between these two alternatives is potentially much greater. Unfortunately, the expectations and variances of crucial quantities (such as tree length) are not yet available for serially sampled data, so this potential power has not been tapped.

Apart from analyses of intra-population sequence variation, evidence for non-neutrality can also be detected by comparing within- and between-species sequence variation [24]. For example, it has been widely observed that in some species there is an excess number of polymorphic non-synonymous sites segregating within the species relative to the number of non-synonymous sites with fixed differences between closely related species [19,25]. This effect is consistent with the conclusion that a substantial fraction of non-synonymous mutations are slightly deleterious mutations (SDMs) that often persist as polymorphisms within populations for some time but have a low probability of eventual fixation [26]. However this pattern is not universal. In fact, at least in *Drosophila* the pattern appears to be the reverse [27], possibly implying a prominent role for recurring positive selection [28]. Regardless of the direction of non-neutral evolution this test may suggest, it has been shown that, as with summary statistics of the mutation frequency spectrum, the accuracy of these methods is compromised by the effects of unrecognized historical demographic change [29]. Both within-species and between-species methods rely on the fact that SDMs become increasingly rare relative to neutral mutations at higher frequencies. For example, within a panmictic population, the distribution of SDMs is expected to predominate near the tips of a population genealogy [30], so that SDMs are on average younger than neutral mutations [25]. Thus the older branches (and associated mutations) within a population will tend to consist of relatively fewer SDMs (as purifying selection has had longer to act).

Although a number of researchers have observed non-neutral behaviour of non-synonymous polymorphism in protein-coding regions, few have considered the effect of SDMs on linked genetic variation in non-coding regions. This is particularly pertinent to the study of the control

region of mitochondrial genes, which is extensively used for within-population genetic sampling of animal mtDNA [31]. The action of Hill-Robertson interference is expected to exacerbate the persistence of SDMs in populations [32], because it reduces the efficiency of purifying selection. Even moderately deleterious mutations, which would otherwise be removed by selection very quickly, can persist in the population if there is substantial genetic linkage between sites [30]. Therefore, in non-recombining genetic elements such as the mitochondrial genome and the genomes of negatively stranded RNA viruses, mutations that are themselves selectively neutral will nevertheless tend to share the fate of linked deleterious mutations.

In this paper we extend an existing Bayesian method originally applied to investigating non-neutrality in HIV evolution [33], that can be used to test for selective neutrality in both coding and non-coding genetic regions sampled from within a single population. The method assumes no knowledge of ancestral mutation frequencies and takes into account the confounding effects of demographic history. We demonstrate the utility of this method on four examples comprised of one non-coding data set and three coding data sets. This method assumes a single genealogy describes the evolutionary history of the sequences under study, but makes no assumptions about ancestral mutation frequencies and takes into account the confounding effects of demographic history. We demonstrate the utility of this method on four non-recombinant examples comprised of one non-coding data set and three coding data sets.

Results and discussion

We employed a suite of summary statistics to test the assumption of neutrality on four example data sets. Because selection is expected to change both the distribution of mutations on the tree and the shape of the sample genealogy [30], statistics that measure both of these departures were included in the analysis.

Summary statistics

Fu and Li [2] compared two estimates of population parameter θ that can be derived for a sample of n sequences:

1. the total number of singleton polymorphisms and

2. the total number of segregating sites divided by $a_n = \sum_{k=1}^{n-1} k^{-1}$.

Under neutrality the difference between these two measures is expected to be zero, and the variance in the difference can be calculated. The resulting normalized test statistic D_F assumes an infinite sites model of mutation, because it equates mutations with branch lengths in the underlying coalescent tree and does not therefore account for the possibility of multiple mutations at a single site. To avoid this assumption we employ a genealogy-based version of D_F, which compares the length of terminal branches to the total length of the coalescent genealogy (we term this the *genealogical* D_F). In addition to the genealogical D_F, two other measures of branch length distribution (age of most recent common ancestor, and total tree length; see Table 1) and three measures of tree imbalance B_1, I_c and C_n were also employed.

The B_1 statistic is the maximum number of nodes between an internal node and the tips of the tree, summed over all internal nodes and excluding the root [34]. Higher values of B_1 are expected with increasing symmetry of the phylogeny. Colless's tree imbalance index I_c considers each internal node of a bifurcating tree and partitions the number of terminal sequences that descend from it into two groups, r and s, where $r \geq s$. Symmetry is measured based on the difference between r and s, summed over all internal nodes [18]. The measure increases from 0 for a perfectly symmetrical tree, to 1 if the tree is completely asymmetric. The final tree-asymmetry measure, Cherry count C_n, is simply the number of pairs of sequences joined by their most recent common ancestor [20]. More symmetrical trees are identified by higher values of C_n. All six summary statistics used are listed in Table 1.

Table 1: Summary statistics used in test of neutrality

Summary Statistic	Reference	Description
T	-	The total length of all branches of the tree.
t_{MRCA}	-	The difference in age between the most recent common ancestor and the most modern individual.
D_F	[2]	A classic summary statistic for testing neutrality. Normalized difference between external branch lengths and total tree length.
B_1	[34]	A measure of tree-imbalance.
C_n	[20]	The number of internal nodes with exactly two terminal children(the number of cherries).
I_c	[18]	A measure of tree-imbalance. Ranges is [0,1]. Larger numbers signify more imbalanced trees.

Data analysis

Brown bear mitochondrial DNA

An alignment of non-coding mitochondrial DNA from the d-loop of brown bears *Ursus ursus* was compiled as an example of non-coding molecular sequence data that is assumed to be evolving neutrally. The data set comprised 30 previously published ancient DNA sequences [35], along with 44 modern brown bear sequences obtained from GenBank. The software BEAST [36] was used to conduct Bayesian MCMC analysis on the full data set (n = 74), yielding estimates of evolutionary rate, population size and ancestral genealogy (Table 2). The substitution model chosen allowed for different rates of transitions and transversions [37] as well as Γ-distributed rate heterogeneity among sites [38]. Both constant-size and exponential-growth models of demography were investigated. To test if the assumption of neutrality was warranted, posterior and posterior predictive values were calculated for each of the summary statistics in Table 1, along with their corresponding multivariate posterior predictive *p*-value. A Bayes factor computed via importance sampling [39,40] was used as a model choice criterion to compare the relative marginal likelihoods of the two models, resulting in rejection of the exponential growth model in favour of a constant-size population. However under both models, differences between the posterior and posterior predictive values did not suggest any significant departures from neutrality in any of the six summary statistics investigated. The multivariate *p*-values for constant and exponential growth were 0.219 and 0.284 respectively. This result suggests, at least in terms of tree asymmetry and branch length distribution, that selective neutrality cannot be rejected for the d-loop of brown bears.

RNA virus data sets

Three RNA virus data sets were also analyzed under the same model conditions as described above. The first was a multiple sequence alignment (n = 129) of the g gene (L = 629 bp) of human respiratory syncytial virus (HRSV) spanning 46 years from 1956 to 2002 [41]. This virus was used as an example of a coding gene of an RNA virus that exhibits only a weak signal of non-neutrality in terms of its tree shape. The estimates of mutation rate and population size are shown in Table 2. A constant population size was preferred over exponential growth using a Bayes factor. The multivariate posterior predictive *p*-values did not reject neutrality (p = 0.33). We followed up with a series of univariate analyses using the individual summary statistics. The tree length T, age of the root t_{MRCA} and D_F statistics are all close to significance under the assumption of constant population size, as shown in Table 3, while the remaining univariate statistics are less suggestive. Therefore, there is only marginal evidence for low levels of non-neutrality in the tree shape of HRSV.

To demonstrate the ability of this method to detect non-neutrality, two additional data sets were analyzed. The first was a previously published data set of the E gene of the dengue-4 virus (n = 69, L = 1485) from Puerto Rico [42] spanning 17 years. The second was a data set of hemagglutinin sequences from human influenza A virus selected to have a similar time frame (1981–1998). These two viral data sets are both expected to exhibit the effects of adaptive selection, particularly influenza A virus, given the nature of their life histories [42-44]. As for the previous data sets, posterior and posterior predictive values were calculated for each of the summary statistics in Table 1. Under constant population size, the multivariate posterior predictive *p*-value = 0.0269 for the dengue-4 virus data set and = 0.0240 for the influenza A virus data set.

Table 2: Bayesian parameter estimates

Data Set	Demographic Model	log $P(D\|M)$[1]	$N_e\tau$ (years)	r[2]	μ[3]	α[4]	κ	t_{MRCA} (years)
Brown bear (d-loop)	Constant*	-2200	113,800	-	5.68×10^{-7}	0.243	41.8	153,500
	Exp. growth	-2198	127,000	5.45×10^{-6}	5.95×10^{-7}	0.243	41.7	145,100
HRSV (g gene)	Constant*	-6068	36.3	-	0.00242	0.900	12.4	56.1
	Exp. growth	-6070	53.0	0.0263	0.00239	0.900	12.4	55.8
Dengue-4 (E gene)	Constant	-3960	11.2	-	0.000976	0.167	17.3	19.7
	Exp. growth*	-3952	38.9	0.134	0.00096	0.167	17.2	19.0
Influenza A (HA)	Constant*	-4386	4.3	-	0.00503	0.332	5.49	19.0
	Exp. growth	-4383	7.25	0.0681	0.00506	0.332	5.5	18.9

Posterior parameter estimates from the MCMC analyses. The effective population size is reported only as a product with generation time ($N_e\tau$) and the compound parameter has unit of years for virus data sets and radiocarbon years for the brown bear data set. Posterior means are reported for all model parameters. For each data set, the demographic model chosen by a Bayes factor is marked (*). [1]marginal likelihood, [2]exponential growth rate, [3]substitution rate, [4]shape parameter of the Γ-distribution.

Table 3: Predictive Probabilities

Data Set	Demographic Model	T	t_{MRCA}	D_F	I_c	C_n	B_1	MV
Brown bear (d-loop)	Constant	0.205	0.164	0.128	0.307	0.844	0.900	0.219
	Exponential growth	0.382	0.374	0.209	0.305	0.832	0.886	0.284
HRSV (g gene)	Constant	0.045	0.034	0.044	0.835	0.851	0.865	0.330
	Exponential growth	0.294	0.335	0.121	0.805	0.845	0.857	0.463
Dengue-4 (E gene)	Constant	0.036	0.004*	0.001*	0.434	0.401	0.581	0.027*
	Exponential growth	0.219	0.170	**0.013***	0.449	0.349	0.498	0.128
Human influenza A (HA)	Constant	0.040	0.101	**<0.001***	0.951	0.392	0.393	**0.024***
	Exponential growth	0.085	0.381	0.001*	0.916	0.427	0.438	0.018*

Univariate and multivariate posterior predictive *p*-values for summary statistics on each of the example data sets. Significant departures (univariate: $p_B < \alpha/2$ or $p_B > 1 - \alpha/2$; multivariate: $p_B < \alpha$ for $\alpha = 0.05$) from neutrality are marked (*). Significant departures on the best fitting model for each data set are in bold.

Both of these data sets exhibited significantly more negative D_F then expected under neutrality, suggesting that the relative length of the terminal branches is larger than expected in both data sets. Additionally, the human influenza A data set also had marginally more tree-imbalance than expected under neutrality, and for the dengue-4 data set, the age of the most recent common ancestor was significantly smaller than expected under neutrality. Figure 1 shows the posterior and predictive distributions for D_F and tree length T for all four data sets. Figure 2 shows (A) two human influenza A virus trees from the posterior distribution of the exponential growth analysis along with (B) the corresponding simulated trees in the predictive distribution. These observations provide qualitative evidence for the ability to detect non-neutral evolutionary dynamics from tree shape as suggested in a recent review of the nascent field of phylodynamics [16].

Distinguishing selection from exponential growth

A criticism often leveled at tests of neutrality such as D_F, is that significantly negative values of D_F could signify exponential growth rather than non-neutral evolution. As demonstrated above the methodology employed here allows the demographic history to be described parametrically as part of the model. Therefore, inference and testing can both be achieved under a model of exponential growth. In this case, any additional departure from expectations cannot be attributed to exponential growth as the demographic signal is incorporated into the test via the predictive distribution. The results of the tests including exponential growth are also presented in Table 3. Interestingly, model selection by Bayes factors can not strongly reject constant populations in all of the data sets except for dengue-4. In the case of dengue-4, the log Bayes factor in favor of the exponential population model is approximately 8. However, the multivariate *p*-value for dengue-4 is no longer significant once exponential growth is incorporated. We can therefore distinguish between selection and growth in the dengue-4 and influenza data sets. In dengue-4, the departure from neutral expectations can be explained by an incorrect choice of demographic functions. Whereas in influenza, significant departures from neutral expectations are observed under both demographic scenarios. In contrast, there is little evidence of non-neutrality in the bear and HRSV data sets.

Simulations

For infinitely long sequences, for which no uncertainty in the underlying genealogy exists, p_B behaves like a classical *p*-value. In the infinite data situation, the posterior distribution of $\mathbf{T}(\cdot)$ collapses to a single point and equation (5) then returns the probability of observing a test statistic under the null hypothesis of selective neutrality as extreme as the test statistic of the data. In finite data situations, p_B is stochastically less variable than a uniform distribution but with the same mean. This implies that the distribution of p_B is more centered about 1/2 than a uniform random variable, leading to slightly more conservative tests when one choses small Type I error rates. To test the assertion that p_B can still be interpreted as a *p*-value even when sequences are of short length and there is significant uncertainty in the underlying genealogy, a simulation study was undertaken. A number of replicate data sets ($n = 100$) were simulated and analyzed as follows:

1. A time-structured coalescent tree was simulated with sample times at 0, 300, 600, and 900 days, with 10 sequences at each time and a constant population-size parameter $N_e\tau = 1500$ (the product of N_e and generation length in days).

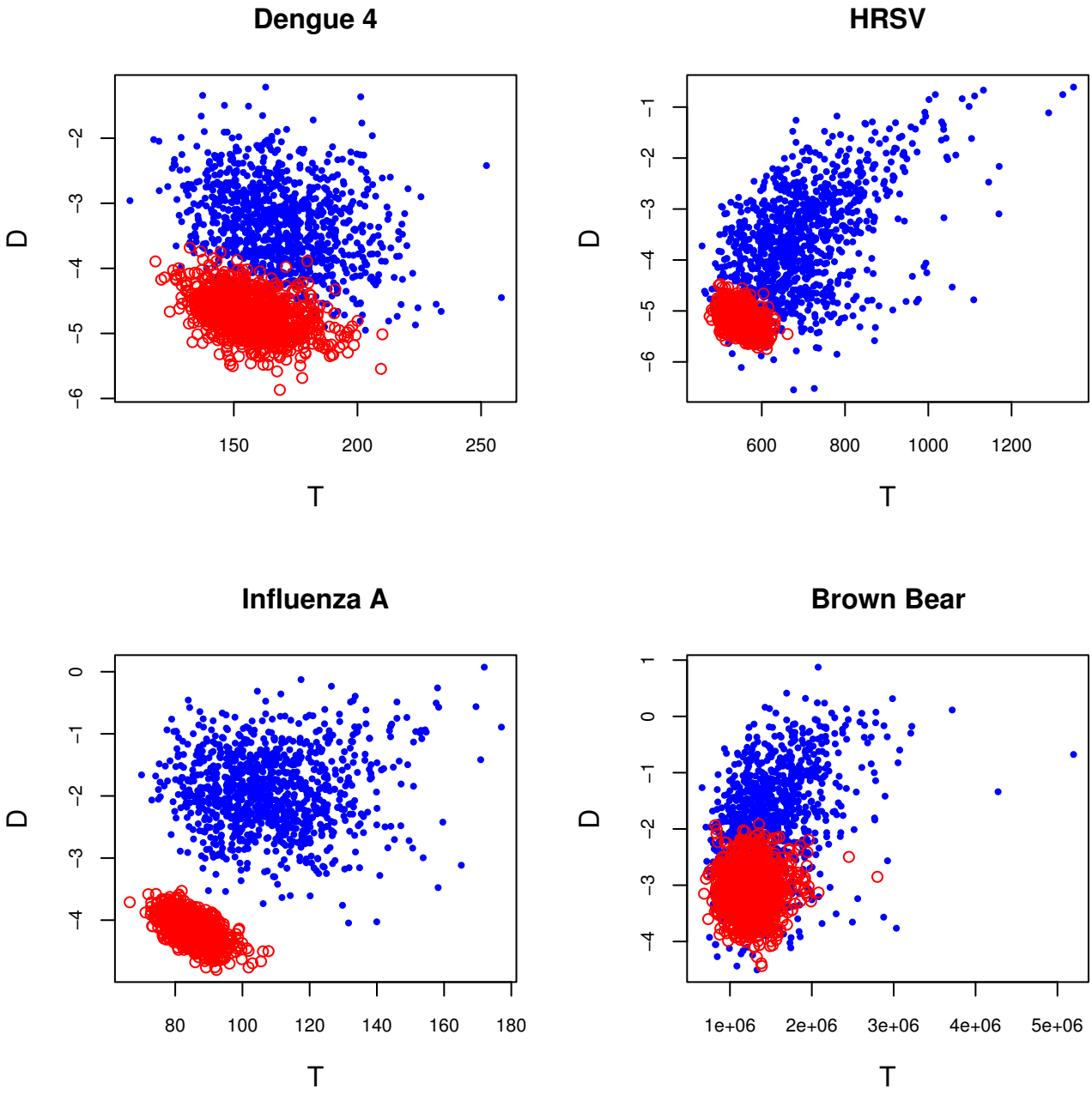

Figure 1
Posterior and predictive distributions of tree length *T* and D_F. Posterior and predictive distributions of tree length *T* and D_F for all four data sets. The dengue-4 data is from an analysis assuming exponential growth, while the other three analyses assumed a constant population size. Human influenza A virus shows the largest departure from neutrality, with the posterior distribution completely disjoint from the predictive distribution.

2. DNA sequences of length 400 were simulated down the coalescent tree under an HKY85 + Γ model of substitution with parameters $\kappa = 8$, $\sigma = 0.1$, and $\mu = 4.0 \times 10^{-5}$ per site per day. Insertions and deletions were not simulated.

3. A Bayesian MCMC analysis was run on the resulting DNA sequence alignments using BEAST (Drummond and Rambaut 2004), assuming a constant population and an HKY85 + Γ model of substitution. The demographic and

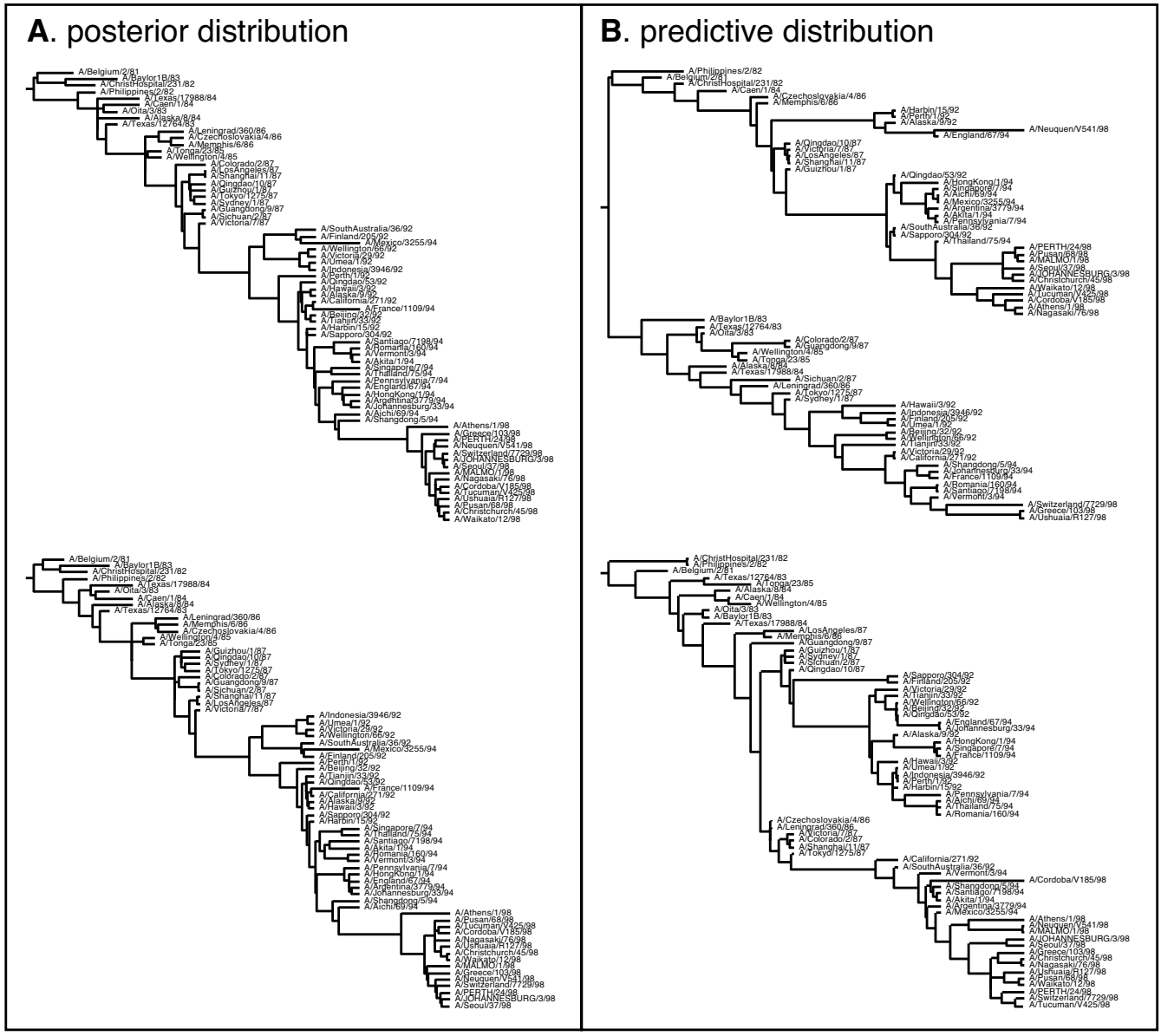

Figure 2

Posterior and posterior predictive genealogies of human influenza A virus. (A) A sample of two trees from the posterior distribution of the human influenza A virus data set. (B) The two matching trees simulated for the predictive distribution of the human influenza A virus data set. Obvious differences between the posterior and predictive trees are the shorter tree length and absence of deep splits in the posterior trees.

substitution parameters were all estimated assuming flat priors with conservative upper bounds.

4. For each state p of the MCMC, the sampled population size parameter $\theta^{(p)} \sim P(\cup|\mathbf{Y})$ was used to generate a time-structure coalescent tree $G^{\text{rep},(p)}$. The set of trees for p = 1, ..., P is the predictive distribution of genealogies.

5. Using equation (5), the posterior distribution of genealogies was compared with the predictive distribution of genealogies, resulting in a p_B value (using the D_F statistic to summary the genealogies, as D_F proved most powerful on the real data sets).

In the above scheme, the model used to simulate the data is the same as the model that we are testing against. Therefore we would expect the p_B values to be distributed approximately uniformly between 0 and 1 under the null hypothesis. Figure 3 shows the cumulative probability distribution for the p_B statistics calculated using the above

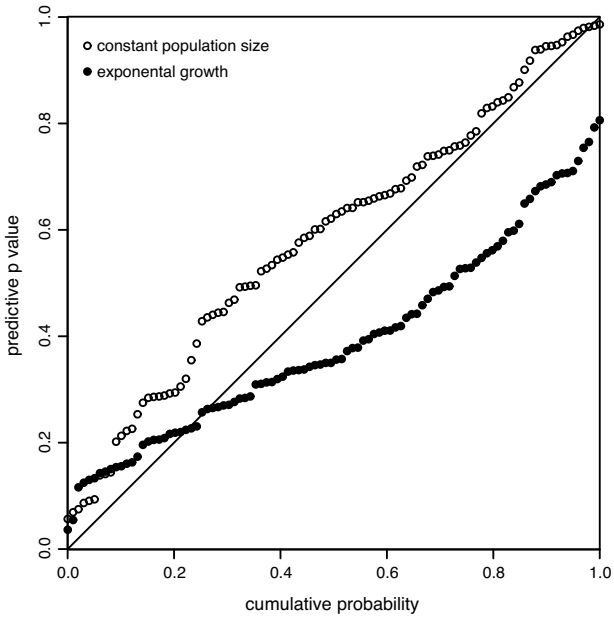

Figure 3
Cumulative distribution of p_B. (Cumulative distribution of p_B values (based on D_F statistic) on 100 simulated data sets under a constant population (open circles) and an exponentially growing population (closed circles). The ideal behaviour for p_B when applied to data simulated from the null distribution would be a uniform distribution (see main text for details). This plot shows that if the true demographic history is a constant population, then p_B will be a good test of neutrality. However, if the true demographic history is exponential growth p_B will be a conservative test, as can be seen by the lack of high p_B values in the closed circles.

scheme as well as for a 100 replicates where exponential growth ($N_e\tau = 5000$, $r = 2 \times 10^{-3}$) was assumed instead of a constant population size. For the case of constant population size, the p_B values are distributed approximately uniformly, with 8 false positives (i.e. $p_B < 0.05$) suggesting that the test is neither overly sensitive nor too conservative (Figure 3). However for the case of exponential growth the values of p_B do not appear uniform with too few extreme values suggesting that the test would be conservative. These results strengthen the conclusion of a relative abundance of external-branch mutations in the viral data sets analyzed in this paper, because the significant statistics observed for real data sets both under constant-size and exponential growth assumptions would not be expected under neutrality.

Conclusion

The results presented here demonstrate the utility of posterior predictive simulation for testing the goodness-of-fit of population genetic models of molecular evolution. In particular we tested the assumption of neutrality under both constant population size and exponential growth on four example data sets where temporally spaced data was available. In both dengue-4 and the human influenza A viruses there was a significant excess of mutations on terminal branches whether or not exponential growth was assumed. In contrast gene trees of HRSV and the d-loop of brown bears did not exhibit any significant departure from neutral expectations in terms of tree shape or genealogical distribution of mutations, although all four data sets had greater than average numbers of mutations on terminal branches relative to internal branches when compared to neutral expectations. Furthermore all four data sets had below average age of the root and below average tree length. In terms of tree-imbalance, both above and below average imbalances were observed for all three tree-imbalance statistics measured (B_1, C_n, I_c).

This paper has been primarily concerned with demonstrating the utility of using existing summary statistics for testing neutrality in temporally spaced data sets. While we have demonstrated that existing statistics, such as D_F can be successfully used to uncover non-neutral evolution it remains likely that better summary statistics may exist. We have described a method for comparing measures of tree shape with their expectations even if the tree shape statistic cannot be directly calculated from the sequence data. We hope that further development of test statistics of tree shape explicitly designed for temporally spaced data will proceed. By doing this we hope tests of recent phylodynamic theories [16] of genetic diversity and evolution in viral pathogens can be constructed. With the posterior predictive framework outlined here, new statistics should greatly increase our ability to detect non-neutral evolution and other departures from standard models of molecular evolution and population genetics. One potentially fruitful direction lies in examining violations of neutrality in the underlying substitution process, as well as in tree-shape. Efficient methods to detect substitution model violations by comparing the expected numbers of different classes of nucleotide substitutions have already been introduced [45]. This allows future work to combine appropriate summary statistics across the full model parameter space in order to maximize statistical power to detect non-neutral evolution.

Our reliance on posterior predictive simulation may raise the concern that the observed data **Y** for each example is "used" twice, first in generating the posterior distribution of model parameters and then in estimating the test-statistic employed to reject the null hypothesis. An alternative approach utilizing prior predictive simulation exists [46] and satisfies the above criticism. However, prior predictive simulation is undefined under improper prior distributions [47] and may not offer sufficient statistical power when vague priors are employed [48,49], such is the case

in this work. In general, if the model parameters G, Ω and θ are well-estimated given **Y**, then the posterior predictive p-value yields results similar to classical p-values (when available), while the prior predictive assessment is highly sensitive to prior distribution choice [47]. In addition to these fully Bayesian predictive methods, Bayes factors [50] are effective model selection tools in phylogenetics [51]. A Bayes factor measures the relative likelihood of two competing models. To compute the Bayes factor in favor of the null model of neutrality, one must specify an alternative model. Unless the researcher has firm *a priori* knowledge about how neutrality might be violated in their data, we recommend starting with rejecting the null through these predictive methods and only then attempting the difficult task of non-neutral model construction and fitting.

Although both dengue-4 and the human influenza A viruses exhibit very ladder-like trees that are highly imbalanced, our analysis suggests that this amount of imbalance is not much more than would be expected given the sampling scheme and estimated effective population sizes. However it can be argued that small effective population sizes, by themselves, are evidence for selection. This is because effective population size (N_e) is a measure of the number of productively replicating individuals, only when the population is evolving under conditions of neutrality. In the absence of any such prior assumptions, N_e should be considered only as a surrogate measure of diversity in the population. Because diversity is reduced by selection, a low estimated N_e could be a sign that the population process is being driven by natural selection. Nevertheless, the results presented here emphasize that ladder-like trees, by themselves, do not necessarily suggest selection. Consequently, interpretation of tree shape imbalance should not be made in the absence of an understanding of the expectations under the null model. Overall, for the example data sets chosen in this study, tree shape did not seem to be a powerful indicator of non-neutral evolution. Finally, by incorporating a demographic model into the test framework, we have ruled out exponential growth as the reason for significant predictive probabilities (p_B) in all data sets besides dengue-4. Nevertheless there remain a number of alternative explanations for neutrality being rejected.

Both human influenza A and dengue-4 viruses show a significant excess of mutations on terminal branches when compared to the predictions of the best fitting parameters of the neutral model. These departures from neutrality lend insight into the process of molecular evolution in RNA viruses, and suggest that new models that take into account these departures need to be developed to accurately model their genetic variation. In contrast, at least with respect to tree shape and genealogical distribution of mutations, neutrality seems to be an approximately adequate model for the G gene of HRSV and the d-loop of brown bears. We hope that further application of posterior predictive simulation will shed light on the pattern of within-population genetic variation in a wide range of species and genetic elements.

Methods

To assess selective neutrality in evolution, traditional test statistics summarize either the observed sequence data **Y** directly or the shape and inter-node distribution of a fixed gene genealogy G relating the sequences, where G is assumed known. In general, however, G is unknown *a priori* and must also be inferred from the sequence data with considerable uncertainty for measurably evolving populations [23]. This presents a difficulty for classical statistical tests. We overcome this short-fall in a Bayesian framework using posterior predictive assessment of model fit [33,47]. In this framework, we estimate G and its associated uncertainty from **Y** using a statistical model of molecular evolution and population demography and simultaneously compare a summary statistic of the random genealogy G to the statistic's expectation under neutrality. Our approach relies on assuming a statistical model for molecular evolution under neutrality. We employ a standard choice based on a continuous-time Markov chain process for nucleotide substitution [52] and an underlying coalescent process to generate the genealogy [53]. In particular, we assume the [37] (HKY85) substitution model with discrete – distributed rate heterogeneity across sites [38] parameterized by $\Omega = (\mu, \kappa, \sigma)$. Parameter μ is the overall rate of mutation, κ is the transition/transversion bias and σ is the Gamma shape parameter. We assume a demographic coalescent process that allows for exponential population growth parameterized by $\theta = (N_e\tau, r)$. Parameter $N_e\tau$ is the product of the effective population size and generation time and r is the exponential growth rate. Restricting $r = 0$ results in a constant population-size model. After assuming a prior distribution over (Ω, θ), we can approximate the posterior distribution

$$P(G, \Omega, \theta|\mathbf{Y}) \quad (1)$$

using Markov chain Monte Carlo (MCMC) techniques [54,55]. We refer interested readers to [22] for further details on prior choices and our MCMC approach. Simulation of (1) is readily available using the software BEAST [36].

With the tools to infer the random genealogy G and model parameters given sequence data in hand, we now consider summary statistics to assess the neutral model fit. Consider a vector of test statistics $\mathbf{T}(G) = [T_1(G), ..., T_K(G)]$ that summarize the shape of the genealogy G. Each element $T_k(G)$ for $k = 1, ..., K$ serves as a unique mapping

between G and the real numbers and generally returns a small value if G were generated by a neutral process and a large value otherwise. One such example for $T_k(G)$ is D_F. Different $T_k(G)$ serve to detect different types of departures from the neutral tree form.

It is important to note that $\mathbf{T}(G)$ depends on an unknown model parameter in contrast to a classical test statistic that depends only on fixed quantities, such as the observed data $\mathbf{Y}$ or a fixed estimate of the genealogy $\hat{G}$. In the Bayesian literature, test statistics that depend on unknown model parameters (and also sometimes the data directly) are generally referred to as "discrepancy values" [48] to help differentiate them from classical measures. To simplify notation, we continue to refer to $\mathbf{T}(G)$ as a summary statistic with the implicit understanding that it is random and not directly observable. The advantage afforded by leaving $\mathbf{T}(G)$ a random variable is that we are now able to compare the discrepancy between the observed data $\mathbf{Y}$ and the posited neutral model as a whole, instead of between the data and the best fit of the model. To use $\mathbf{T}(G)$ to assess the model fit of neutrality, we consider the following thought experiment. Suppose we randomly simulate under a neutral model a genealogy G^{rep} from a replicated population almost identical to the population yielding the sequence data $\mathbf{Y}$, where both populations share the same unknown demographic parameters θ, number of tips and tip-dates. Then, we compare quantities $\mathbf{T}(G^{\text{rep}})$ and $\mathbf{T}(G)$ given $\mathbf{Y}$. Disparate values signify model misspecification caused by non-neutral evolutionary forces.

We recall that $\mathbf{T}(G^{\text{rep}})$ and $\mathbf{T}(G)$ given $\mathbf{Y}$ are not fixed values, but are random variables represented by probability distributions. As a consequence, we must integrate over all possible realizations weighed by their posterior probabilities to generate a test based on $\mathbf{T}(\cdot)$. This process is called posterior predictive simulation [46-48,56]. Model selection and critique using posterior predictive simulation has had a successful history in phylogenetics [33,49,57-59].

The central distribution that we require is the posterior predictive distribution of the test statistic

$$P[\mathbf{T}(G^{\text{rep}}) \mid \mathbf{Y}] = \int_G \int_{\boldsymbol{\Omega}} \int_{\theta} P[\mathbf{T}(G^{\text{rep}}) \mid \theta] P(G, \boldsymbol{\Omega}, \theta \mid \mathbf{Y}) \mathrm{d}\theta \mathrm{d}\boldsymbol{\Omega} \mathrm{d}G. \quad (2)$$

In practice, one approximates the predictive distribution in (2) by first generating a posterior sample $\{G^{(p)}, \Omega^{(p)}, \theta^{(p)}\}$ for $p = 1, \ldots, P$ from $P(G, \Omega, \theta|\mathbf{Y})$. Then, for each p, one draws

$$G^{\text{rep},(p)} \sim P(\cdot \mid \theta^{(p)}), \quad (3)$$

where $P(G^{\text{rep}}|\theta)$ describes a selectively neutral coalescent process. Finally, one tabulates $\mathbf{T}(G^{\text{rep},(p)})$. We interpret this predictive distribution as a description of the values that $\mathbf{T}(\cdot)$ generates when applied to genealogies from selectively neutral populations. To assess neutrality in the observed data, we compare the predictive distribution to the posterior distribution of the test statistic

$$P[\mathbf{T}(G) \mid \mathbf{Y}] = \int_{\boldsymbol{\Omega}} \int_{\theta} P[\mathbf{T}(G), \boldsymbol{\Omega}, \theta \mid \mathbf{Y} \mid \mathrm{d}\boldsymbol{\Omega} \mathrm{d}\theta, \quad (4)$$

approximated by tabulating $\mathbf{T}(G^{(p)})$ for $p = 1, \ldots, P$.

When the test statistic $\mathbf{T}(\cdot)$ is univariate [33], assessing differences between predictive and posterior distributions can be done in two ways [47]. The first method is graphical, generating a scatterplot of $\{[\mathbf{T}(G^{\text{rep},(p)}, \mathbf{T}(G^{(p)})], p = 1, \ldots, P\}$. The second method is more formal, employing tail-area probabilities.

Let the posterior predictive p-value [48]

$$p_B = P[\mathbf{T}(G^{\text{rep}}) \geq \mathbf{T}(G)|\mathbf{Y}], \quad (5)$$

then p_B remains well-defined even though $\mathbf{T}(G^{\text{rep}})$ and $\mathbf{T}(G)$ given $\mathbf{Y}$ are not directly observable [47]. Probability p_B shares many characteristics with a classical p-value; for example, p_B can be viewed as its posterior mean and, under the null hypothesis of neutrality, p_B is approximately distributed as a Uniform [0, 1) random variable [48]. Given these properties, we reject the selectively neutral model for extreme values of p_B, say $p_B < \alpha = 0.05$ for strictly non-negative $\mathbf{T}(\cdot)$ or $p_B < \alpha/2$ or $p_B > 1 - \alpha/2$ otherwise.

To calculate p_B, a consistent estimator is

$$\hat{p}_B = \sum_{p=1}^{P} 1\{\mathbf{T}(G^{\text{rep},(p)}) \geq \mathbf{T}(G^{(p)})\} \quad (6)$$

where $1\{\cdot\}$ is the indicator function, returning 1 if its argument is true and 0 otherwise.

When the test statistic $\mathbf{T}(\cdot)$ is multivariate, we are able to detect a greater variety of departures from selective neutrality simultaneously, but a single tail-area probability becomes more troublesome to calculate. In this situation, we first standardize individual elements $T_k(\cdot)$ such that var $[T_k(G)|\mathbf{Y}] = 1$ for all k. This places all measures on a common scale. We then generate scatterplots of the multivariate distributions. We agree with [47] in that comparing the posterior and predictive distributions graphically

can provide more information than reporting a single *p*-value. For example, we can identify which components $T_k(\cdot)$ in $\mathbf{T}(\cdot)$ contribute greatest to the discrepancy between the data and a selectively neutral model.

To calculate a tail-area probability in the multivariate setting, we turn to the (squared) Mahalanobis distance in constructing a posterior predictive test [60]. Let $\hat{\mathbf{m}}$ be an estimate of the predictive mean of $\mathbf{T}(G^{\text{rep}})$ and $\hat{\mathbf{V}}$ be an estimate of its variance-covariance matrix, such that

$$\begin{aligned}\hat{\mathbf{m}} &= \frac{1}{P}\sum_{p=1}^{P}\mathbf{T}(G^{\text{rep},(p)}), \quad \text{and} \\ \hat{\mathbf{V}} &= \frac{1}{P-1}\sum_{p=1}^{P}\left[\mathbf{T}(G^{\text{rep}(p)}) - \hat{\mathbf{m}}\right]^t\left[\mathbf{T}(G^{\text{rep},(p)}) - \hat{\mathbf{m}}\right],\end{aligned} \tag{7}$$

for $p = 1, \ldots, P$. Then, we define the (squared) Mahalanobis distance

$$M(\mathbf{x}) = (\mathbf{x} - m)^t\mathbf{V}^{-1}(\mathbf{x} - m), \tag{8}$$

where we substitute $\mathbf{T}(G)$ for $\mathbf{x}$ when considering the distance's posterior distribution and $\mathbf{T}(G^{\text{rep}})$ for $\mathbf{x}$ when considering its predictive distribution. Mahalanobis distances are commonly used in discrimination analysis and classification. The metric of the Mahalanobis distance $M(\cdot)$ is the inverse of the variance-covariance matrix $\hat{\mathbf{V}}$ of the predictive distribution and, as such, returns distances normalized relative to the multidimensional spread of the data under selective neutrality. Following in the light of Equation (5), we define the multivariate posterior predictive *p*-value

$$p_B = P[M(G^{\text{rep}}) \geq M(G)|\mathbf{Y}]. \tag{9}$$

A consistent estimator of the multivariate p_B is readily available in the vain of Equation (6).

When it is unclear *a priori* which elements $T_k(\cdot)$ provide the most power to reject selective neutrality, the multivariate approach side-steps the multiple testing problem inherent in examining each element independently. In these situations, we consider first using (9) as a global test with a fixed Type I Error rate α and then sub-selecting a small number of individual $T_k(\cdot)$ for further univariate analysis. For researchers who begin by examining the *K* univariate analyses separately, we recommend applying a Bonferroni correction by decreasing the critical value cut-off from α to α/K per test. For large *K*, a Bonferrioni correction is overly conservative, especially when considering the potentially high correlation between $T_k(\cdot)$. At this point, monitoring the false discovery rate [61] becomes more practical.

Authors' contributions

AJD conceived the original idea and performed the initial data analysis and wrote the first draft of the paper. MAS constructed and performed the multivariate tests including re-creation of Figures 1 and 3 and Table 3. Both authors contributed to the final text.

Acknowledgements

The DIMACS Working Group on Phylogenetic Trees and Rapidly Evolving Diseases fostered the initial collaboration between A.J.D. and M.A.S. We thank Andrew Rambaut, Eddie Holmes, Oliver G. Pybus and Allen G. Rodrigo for helpful discussions. We thank Charles Edwards and Daniel Wilson for assistance in producing the simulation results. This research was funded in part by Wellcome Trust Grant 017979 (to A.J.D.) and NIH grant GM086887 and the John Simon Guggenheim Memorial Foundation (to M.A.S.).

References

1. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585-595.
2. Fu Y, Li W: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133:**693-709.
3. Fu Y: **Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection.** *Genetics* 1997, **147:**915-925.
4. Fu YX: **New statistical tests of neutrality for DNA samples from a population.** *Genetics* 1996, **143:**557-570.
5. Fay J, Wu C: **Hitchhiking under positive Darwinian selection.** *Genetics* 2000, **155:**1405-1413.
6. Strobeck C: **Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision.** *Genetics* 1987, **117:**149-153.
7. Hudson R, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ: **Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster.** *Genetics* 1994, **136:**1329-1340.
8. Slatkin M, Hudson R: **Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations.** *Genetics* 1991, **129:**555-562.
9. Mousset S, Derome N, Veuille M: **A test of neutrality and constant population size based on the mismatch distribution.** *Molecular Biology and Evolution* 2004, **21:**724-731.
10. Ramos-Onsins S, Rozas J: **Statistical properties of new neutrality tests against population growth.** *Molecular Biology and Evolution* 2002, **19:**2092-2100.
11. Przeworski M: **The Signature of Positive Selection at Randomly Chosen Loci.** *Genetics* 2002, **160:**1179-1189.
12. Haddrill P, Thornton K, Charlesworth B, Andolfatto P: **Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations.** *Genome Research* 2005, **15:**790-799.
13. Innan H, Zhang K, Marjoram P, Tavare S, Rosenberg N: **Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites.** *Genetics* 2005, **169:**1763-1777.
14. Innan H: **Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests.** *Genetics* 2006, **173:**1725-1733.
15. Li H, Stephan W: **Inferring the Demographic History and Rate of Adaptive Substitution in Drosophila.** *PLoS Genetics* 2006, **13:**2-10.
16. Grenfell B, Pybus O, Gog J, Wood J, Daly J, Mumford J, Holmes E: **Unifying the epidemiological and evolutionary dynamics of pathogens.** *Science* 2004, **303:**327-332.

17. Kelly JK: **A test of neutrality based on interlocus associations.** *Genetics* 1997, **146(3):**1197-1206.
18. Colless D: **Review of "Phylogenetics: The Theory and Practice of Phylogenetic Systematics".** *Systematic Zoology* 1982, **31:**100-104.
19. Hasegawa M, Cao Y, Yang Z: **Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species.** *Molecular Biology and Evolution* 1998, **15:**1499-1505.
20. McKenzie A, Steel M: **Distributions of cherries for two models of trees.** *Mathematical Biosciences* 2000, **164:**81-92.
21. Aldous D: **Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today.** *Statistical Science* 2001, **16:**23-34.
22. Drummond A, Nicholls G, Rodrigo A, Solomon W: **Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.** *Genetics* 2002, **161:**1307-1320.
23. Drummond A, Pybus O, Rambaut A, Forsberg R, Rodrigo A: **Measurably evolving populations.** *Trends in Ecology & Evolution* 2003, **18:**481-488.
24. McDonald J, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila.** *Nature* 1991, **351:**652-654.
25. Nielsen R, Weinreich DM: **The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory.** *Genetics* 1999, **153:**497-506.
26. Eyre-Walker A, Keightley P, Smith N, Gaffney D: **Quantifying the slightly deleterious mutation model of molecular evolution.** *Molecular Biology and Evolution* 2002, **19:**2142-2149.
27. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E, Langley CH: **Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans.** *PLoS Biol* 2007, **5(11):**e310.
28. Hahn MW: **Toward a selection theory of molecular evolution.** *Evolution* 2008, **62(2):**255-265.
29. Eyre-Walker A: **Changing effective population size and the McDonald-Kreitman test.** *Genetics* 2002, **162:**2017-2024.
30. Williamson S, Orive M: **The genealogy of a sequence subject to purifying selection at multiple sites.** *Molecular Biology and Evolution* 2002, **19:**1376-1384.
31. Avise J: *Phylogeography: The History and Formation of Species* Cambridge, MA: Harvard University Press; 2000.
32. McVean G, Charlesworth B: **The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation.** *Genetics* 2000, **155:**929-944.
33. Edwards C, Holmes E, Pybus O, Wilson D, Viscidi R, Abrams E, Phillips R, Drummond A: **Evolution of the HIV-1 envelope gene is dominated by purifying selection.** *Genetics* 2006, **174:**1441-1453.
34. Kirkpatrick M, Slatkin M: **Searching for evolutionary patterns in the shape of a phylogenetic tree.** *Evolution* 1993, **47:**1171-1181.
35. Barnes I, Matheus P, Shapiro B, Jensen D, Cooper A: **Dynamics of Pleistocene population extinctions in Beringian brown bears.** *Science* 2002, **295:**2267-2270.
36. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7:**214.
37. Hasegawa M, Kishino H, Yano T: **Dating the human-ape splitting by a molecular clock of mitochondrial DNA.** *Journal of Molecular Evolution* 1985, **22:**160-174.
38. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *Journal of Molecular Evolution* 1994, **39:**306-314.
39. Newton M, Raftery A: **Approximate Bayesian inference with the weighted likelihood bootstrap.** *Journal of the Royal Statistical Society, Series B* 1994, **56:**3-48.
40. Suchard M, Kitchen C, Sinsheimer J, Weiss R: **Hierarchical phylogeneic models for analyzing multipartite sequence data.** *Systematic Biology* 2003, **52:**649-664.
41. Zlateva K, Lemey P, Vandamme A, van Ranst M: **Molecular evolution and circulation patterns of human respiratory syncytial virus subgroup A: positively selected sites in the attachment g glycoprotein.** *Journal of Virology* 2004, **78:**4675-4683.
42. Bennett S, Holmes E, Chirivella M, Rodriguez D, Beltran M, Vorndam V, Gubler D, McMillan W: **Selection-driven evolution of emergent dengue virus.** *Molecular Biology and Evolution* 2003, **20:**1650-1658.
43. Fitch W, Bush RM, Bender C, Cox N: **Long term trends in the evolution of H(3) HA1 human influenza type A.** *Proceedings of the National Academy of Sciences, USA* 1997, **94:**7712-7718.
44. Ferguson N, Galvani A, Bush R: **Ecological and immunological determinants of in uenza evolution.** *Nature* 2003, **422:**428-433.
45. Minin V, Suchard M: **Fast, accuract and simulation-free stochastic mapping.** *Proceedings of the Royal Society, Series B* in press.
46. Box G: **Sampling and Bayes inference in scientific modeling and robustness.** *Journal of the Royal Statistical Society, Series A* 1980, **143:**383-430.
47. Gelman A, Meng X, Stern H: **Posterior predictive assessment of model fitness via realized discrepancies (with discussion).** *Statistica Sinica* 1996, **6:**733-807.
48. Meng XL: **Posterior predictive *p*-values.** *Annals of Statistics* 1994, **22:**1142-1160.
49. Suchard M, Weiss R, Sinsheimer J, Dorman K, Patel M, McCabe E: **Evolutionary similarity among genes.** *Journal of the American Statistical Assocation* 2003, **98:**653-662.
50. Kass R, Raftery A: **Bayes factors.** *Journal of the American Statistical Association* 1995, **90:**773-795.
51. Suchard M, Weiss R, Sinsheimer J: **Bayesian selection of continuous-time Markov chain evolutionary models.** *Molecular Biology and Evolution* 2001, **18:**1001-1013.
52. Lange K: *Mathematical and Statistical Methods for Genetic Analysis* New York, NY: Springer; 1997.
53. Kingman J: **The coalescent.** *Stochastic Processes and their Applications* 1982, **13:**235-248.
54. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E: **Equation of state calculations by fast computing machines.** *Journal of Chemical Physics* 1953, **21:**1087-1092.
55. Hastings W: **Monte Carlo sampling methods using Markov chains and their applications.** *Biometrika* 1970, **57:**97-109.
56. Rubin D: **Bayesianly justifiable and relevant frequency calculations for the applied statisticians.** *Annals of Statistics* 1984, **12:**1151-1172.
57. Suchard M, Weiss R, Dorman K, Patel M, McCabe E, Sinsheimer J: **Evolutionary similarity among genes when data are sparse.** In *Proceedings of the Section on Bayesian Statistical Science* Alexandria, VA: American Statistical Assocation; 2000:92-97.
58. Bollback J: **Bayesian model adequacy and choice in phylogenetics.** *Molecular Biology and Evolution* 2002, **19:**1171-1180.
59. Nielsen R, Huelsenbeck J: **Detecting positively selected amino acid sites using posterior predictive P-values.** *Pacific Symposium on Biocomputing* 2002, **7:**576-588.
60. O'Hagan A: **HSSS model criticism (with discussion).** In *Highly Structured Stochastic Systems* Edited by: Green P, Hjort N, Richardson S. Oxford, UK: Oxford University Press; 2003:423-453.
61. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statisical Society, Series B* 1995, **57:**289-300.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral