Available online at www.sciencedirect.com

ScienceDirect

Infection, Genetics and Evolution 8 (2008) 239–246

Infection, Genetics and Evolution

www.elsevier.com/locate/meegid

Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty

Joe Parker [a,*], Andrew Rambaut [b], Oliver G. Pybus [a]

[a] *Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom*
[b] *Institute for Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom*

Received 8 March 2007; received in revised form 17 July 2007; accepted 3 August 2007
Available online 21 August 2007

Abstract

Many recent studies have sought to quantify the degree to which viral phenotypic characters (such as epidemiological risk group, geographic location, cell tropism, drug resistance state, etc.) are correlated with shared ancestry, as represented by a viral phylogenetic tree. Here, we present a new Bayesian Markov-Chain Monte Carlo approach to the investigation of such phylogeny–trait correlations. This method accounts for uncertainty arising from phylogenetic error and provides a statistical significance test of the null hypothesis that traits are associated randomly with phylogeny tips. We perform extensive simulations to explore and compare the behaviour of three statistics of phylogeny–trait correlation. Finally, we re-analyse two existing published data sets as case studies. Our framework aims to provide an improvement over existing methods for this problem.

Keywords: Virus; Molecular epidemiology; Phylogeny; Parsimony; Correlation; HIV; Dengue

1. Introduction

In recent years, explosions in the availability of molecular sequence data and of statistical methods for evolutionary analysis have given new insights in the field of molecular epidemiology. For example, the processes of natural selection, recombination, mutation and migration have all been studied to great effect at different levels of biological organization. However, despite recent increases in computing power, analytical approaches for some classes of problem are still in need of further improvement and rigorous statistical validation.

One such under-developed area concerns the association of phenotypic characters (e.g. geographic locations, physical characteristics, behavioural traits, etc.) with the shared ancestry of a sample of organisms from which gene sequences have been obtained. The individuals sampled may represent different cells, virions, organisms, populations, species, or even higher phyla. We may wish to know whether a particular phenotype has arisen independently in different organisms, or whether it is the result of common ancestry from a single ancestral individual. Another common application of phylogeny–trait correlation is the investigation of spatial population structure; that is, do sequences cluster on a phylogeny according to their geographic location (e.g. Avise, 2000; Starkman et al., 2003; Holmes, 2004)? In all such cases, analyses are complicated by the lack of statistical independence – the phenotypic traits associated with each phylogenetic tip may not be independent as a result of the shared ancestry among sampled individuals (Harvey and Pagel, 1991). It is therefore inappropriate to use standard general linear models to statistically test the null hypothesis that the phenotypes are uncorrelated with the genetic distances among sampled individuals.

A number of previous studies in the field of molecular epidemiology have investigated the association between a virus phylogeny and viral traits. These have included investigations of population structure, resulting from geographic location (e.g. Cochrane et al., 2002; Nakano et al., 2004; Carrington et al., 2005), epidemiological risk group (e.g. Holmes et al., 1995; Leigh Brown et al., 1997) or compartmentalization, either among different host tissues (e.g. Salemi et al., 2005; Pillai et al., 2006) or among different host cell types (e.g. Fulcher et al., 2004). The detection of within-host compartmentalization has been an issue of particular interest for the human immunodeficiency virus (HIV) (McGrath et al., 2001; Kemal

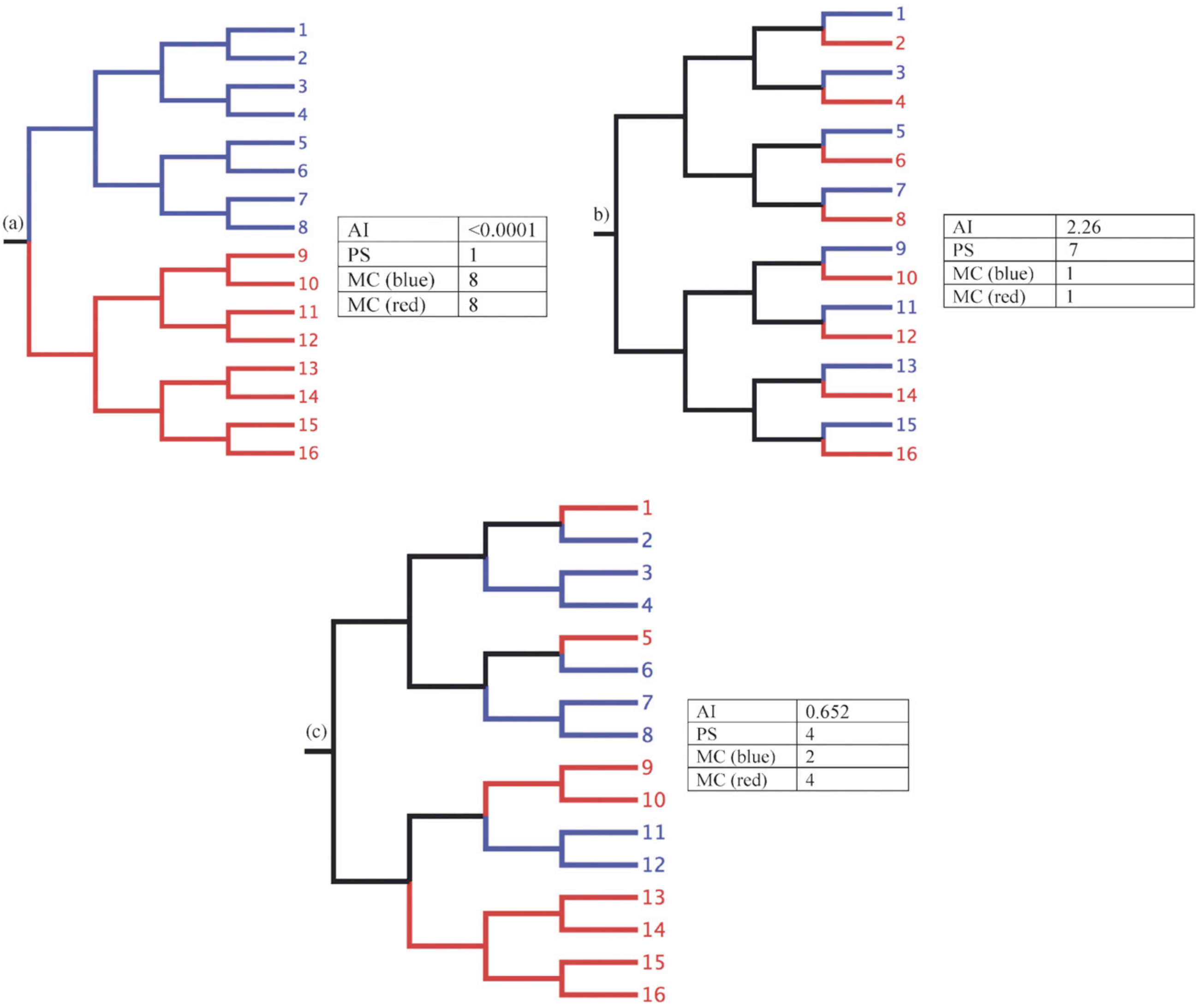
* Corresponding author. Tel.: +44 1865 281987; fax: +44 1865 271249.
E-mail address: joe.parker@zoo.ox.ac.uk (J. Parker).

1567-1348/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2007.08.001

et al., 2003; Sanjuan et al., 2004). In addition, phylogeny–trait associations have been used to investigate antibody and T-cell escape during chronic hepatitis C virus (Sheridan et al., 2004; Komatsu et al., 2006) and HIV (Bhattacharya et al., 2007) infection.

A closely related and long-standing technique is the estimation of gene flow among subpopulations using explicit population genetic models (e.g. Wright, 1952; Beerli and Felsenstein, 2001). Such approaches have recently been implemented in a Bayesian framework and applied to virus genetic data (Beerli and Felsenstein, 2001; Wilson and Rannala, 2003; Ewing et al., 2004). However, given the complexity of such methods it may be desirable to first demonstrate that the sequences are indeed phylogenetically structured according to the trait of interest. In this paper, we investigate and develop statistical tests for this preliminary null hypothesis.

What exactly do we mean by phylogeny–trait correlation? Given a discrete character for each tip of a phylogenetic tree, we are asking if more closely related taxa are more likely to share the same trait values than we would expect by chance alone, i.e. if the characters were randomly assigned to the phylogeny tips. As illustrated in Fig. 1, the tip characters may be tightly correlated with phylogeny (Fig. 1a) or they may be fully interspersed (Fig. 1b). The biological significance of either situation will depend on the nature of the phenotypic trait under investigation. If, for example, the trait represents geographic location then a tight correlation reflects low lineage dispersal or migration, and the opposite represents panmixis or high rates of gene flow. Alternatively, if the trait is thought to be under strong selection – pathogen drug resistance, for example – then interspersed traits may indicate that drug resistance has independently evolved several times or that this phenotype is



AI	<0.0001
PS	1
MC (blue)	8
MC (red)	8

AI	2.26
PS	7
MC (blue)	1
MC (red)	1

AI	0.652
PS	4
MC (blue)	2
MC (red)	4

Fig. 1. (a) *Strong association*: There is clear association between the characters at the tips (represented by taxon colours) and the phylogeny. Taxa 1–8 have inherited the ‘blue’ character state, whilst taxa 9–16 have inherited ‘red.’ (b) *Maximally interspersed*: This tree clearly shows no clear association between phenotype and phylogeny. The ‘blue’ or ‘red’ characters have been acquired or lost multiple times in the ancestry of these taxa. (c) *Intermediate situation*: In some areas of the tree, such as taxa 13–16, it does appear that sister taxa share characters of interest. However, in other areas, such as sister taxa 1–2 and 5–6, the trait values look more interspersed. An analytical method is needed to decide if the association between tips and characters is significant.

not under strong evolutionary constraints. However, in many situations the phylogenetic distribution of the traits may be intermediate and their correlation with phylogeny less clear (Fig. 1c). Therefore, the strength of the association needs to be quantified and statistically tested against the distribution of characters expected by chance.

A variety of metrics have been proposed to quantify phylogeny–trait correlation. An early approach by Hudson et al. (1992) used a range of sequence-summary statistics calculated directly from a sequence alignment. Unfortunately such techniques suffer from a lack of independence due to shared ancestry, as explained above, and as a result more recent techniques have tended to employ some form of phylogeny, either estimated from a molecular sequence alignment or from morphological information.

The most common phylogenetic method is to use a parsimony approach (such as the Fitch, 1971b algorithm) to reconstruct the character states at ancestral nodes. The number of state changes in the phylogeny is then calculated (the parsimony score statistic; PS). However, although PS quantifies phylogeny–trait correlation, it provides no information on whether the value obtained is statistically significant or not. Slatkin and Maddison (1989) addressed this problem by randomizing tip-character associations a large number of times and calculating the PS statistic from each randomization, thereby providing a null distribution of the PS statistic with critical values at the $p = 0.05$ confidence level, against which the observed PS value can be compared. A second metric is the association index (AI) statistic, which explicitly takes into account the shape of the phylogeny by measuring the imbalance of internal phylogeny nodes (Wang et al., 2001; see Section 2). As with the PS statistic, a randomization approach can be used to generate a null distribution for the AI statistic (Wang et al., 2001).

The methods outlined above involve calculating metrics from a single phylogeny that is assumed to be correct. In reality, this single tree is estimated from gene sequences with phylogenetic error, and there may be a large set of different trees that do not differ significantly from the maximum likelihood (ML) tree (Jermiin et al., 1997). Single-tree approaches do not incorporate this error and thus underestimate the true statistical variance. Wang et al. (2001) attempted to address this issue by calculating the AI statistic across a set of bootstrap replicate trees.

In this paper, we concentrate on the methods most commonly applied to viral phylogenies, namely the PS and AI statistics. However, we note that several related statistics have been developed in the field of community ecology. Faith (1992) defined the ‘phylogenetic diversity’ (PD) of a set of taxa as the sum of the shortest paths between all taxa in the set. Webb (2000) and Webb et al. (2002) developed two related methods, the net relatedness index (NRI) and nearest taxa index (NTI), which combine nodal distances (the number of nodes between two taxa that share a trait value) with branch lengths. Lastly, Lozupone and Knight (2005) introduced the UniFrac statistic, which measures the proportion of phylogeny branch lengths that can be unambiguously associated with a particular trait value. A common feature of the PD, NTI/NRI and UniFrac statistics is that they all depend on both the tree topology and the tree branch lengths. The PS and AI statistics, in contrast, depend only on the former. Both types of statistic measure the degree to which taxa with the same trait values cluster together, but the PD, NTI/NRI and UniFrac statistics also measure the genetic similarity among clustering taxa.

In this study, we accommodate phylogenetic error in the calculation of phylogeny–trait correlations using Bayesian Markov chain Monte Carlo (MCMC) methods. Such methods have become increasingly popular and practical over recent years (Holder and Lewis, 2003). Programs that implement MCMC sampling can be used to obtain a posterior distribution of phylogenies, from which the posterior distributions of phylogeny shape statistics can be calculated. Our method correctly incorporates statistical error arising from phylogenetic uncertainty and provides error intervals for hypothesis testing, as well as returning the posterior distributions of the statistics, which provides greater detail than the traditional single ‘p-value’. Additionally, we perform extensive simulations to test the statistical behaviour of different statistics for the first time. We also investigate a new phylogeny–trait statistic, the ‘MC size’ statistic (described below). Finally, we investigate the performance of our new method by re-analysing the data published in Carrington et al. (2005) and Salemi et al. (2005); these data were previously investigated using other phylogeny–trait correlation methods.

2. Methods

We start by defining how phylogeny–trait statistics are calculated from a single phylogeny. Fig. 1 provides the values of several different statistics for three example trees. We then explain how phylogenetic uncertainty can be incorporated into this calculation.

The parsimony score (PS) statistic can be calculated using the Fitch (1971b) parsimony algorithm. If the gain/loss of the trait under investigation does occur parsimoniously, then the observed PS value should be inversely related to the strength of tip-character association. The PS statistic for a given trait takes the range $1 \leq \mathrm{PS} \leq n$, where n is the number of tips in the phylogeny. Low PS scores represent strong phylogeny–trait association. Note that for a single tree, PS (unlike AI) takes integer values and hence is a discrete metric.

The association index statistic introduced by Wang et al. (2001) is the sum:

$$\mathrm{AI} = \sum_{i=1}^{k} \frac{1 - f_i}{2^{m_i - 1}} \qquad (1)$$

The AI is a sum across all the internal nodes in the phylogeny; k is the number of internal nodes. For each internal node i, f_i is defined as the frequency of the most common trait value among the tips subtended by that node; m_i is the number of tips subtended by node i. Thus, low AI values represent strong phylogeny–trait association.

We also define a new statistic that was used in Salemi et al. (2005) but which has not been previously investigated.

Intuitively, stronger phylogeny–trait associations should produce larger monophyletic clades whose tips all share the same trait. This property is quantified by the monophyletic clade (‘MC’) size statistic for a particular trait value x, *defined* as:

$$\mathrm{MC}(x) = \max_{i=1}^{k}(m_i I_i) \quad (2)$$

where m_i is the number of tips subtended by node i and I_i is an indicator function that equals 1 if all tips subtended by node i have trait value x, and equals zero otherwise. k is the number of internal nodes in the phylogeny, including the root. MC is a discrete integer metric for a single tree and is bounded by $1 \leq \mathrm{MC} \leq n_x$, where n_x is the number of tips that have trait value x. MC will be positively correlated with the strength of the phylogeny–trait association.

2.1. Incorporating phylogenetic error

The above methods all require a fully resolved phylogeny to be specified *a priori*. In practice, the tree is estimated from sequences and has an associated statistical error. To account for phylogenetic uncertainty, we developed a Bayesian MCMC approach. Programs such as BEAST (Drummond et al., 2002; Drummond and Rambaut, 2003) or MrBayes (Huelsenbeck and Ronquist, 2001) calculate a posterior sample of trees (PST) that approximates the true posterior distribution of phylogenies given the sequences, with more likely phylogenies being sampled more frequently, and less likely ones less so. By calculating and averaging phylogeny–trait statistics across all trees in the posterior sample, we integrate over (marginalize) the phylogeny and thus incorporate phylogenetic uncertainty.

We combine phylogenetic error and significance testing in the following way. First, the value of the statistic concerned is calculated for every tree in the posterior sample, forming the posterior distribution of the statistic, and the median of this posterior distribution is denoted μ. Next, from the observed set of taxon–character associations C, we generate n randomized sets of taxon–character associations $\{C_1, C_2, C_3, \ldots, C_n\}$. Each randomized set C_i is simply the observed set of associations C resampled without replacement. The set $\{C_1, C_2, C_3, \ldots, C_n\}$ therefore constitutes a null distribution of taxon–character associations; in our analyses we used $n = 100$. Then, for each C_i the median posterior estimate of the statistic is calculated from the PST using the same method as for the observed data, and denoted μ_i. The distribution of the μ_i obtained from the n randomized sets therefore corresponds to an estimate of the null distribution of the statistic. The significance p is then obtained from this null distribution by simply calculating the proportion of μ_i values that are more extreme than the observed value μ (low values being extreme for AI and PS; high values being extreme for MC).

We have developed a computer program BaTS (Bayesian Tip-association Significance testing; available on request) that takes the PST output from an existing program such as BEAST or MrBayes and performs these randomizations.

3. Simulations

We performed a set of simulations to test the type 1 statistical error of our Bayesian MCMC approach (i.e. the probability of rejecting a null hypothesis when the null hypothesis is true). Theoretically, if data are repeatedly simulated under the null hypothesis, then the distribution of the resulting p-values should follow a unit uniform distribution. If this is so, then the type 1 error of the test will correct for all levels of statistical significance (e.g. $p = 0.05$, 0.01, etc.). We therefore simulated a large number of data sets with random tip–trait associations and computed the p-values for each test statistic on each dataset. The p-values were then collated to create a cumulative density function, which was compared against the expected unit uniform distribution using the two-sample Kolmogorov–Smirnoff test.

Data sets were simulated in two steps: (i) a large set of simulated alignments were generated and (ii) character traits were randomly associated with the simulated sequences. For the first step, 3436 random trees of 32 taxa were generated under a pure-birth process using the package Phyl-O-Gen (Rambaut, 2001). This set therefore includes phylogenies with a wide variety of tree shapes, branch lengths and node imbalances. Next, the Seq-Gen (Rambaut and Grassly, 1997) program was used to create sequence alignments, by simulating down each tree using a substitution model typical of empirical HIV-1 *env* gene data sets (transition/transversion ratio = 2.4; base frequencies $A = 0.426$, $C = 0.152$, $G = 0.182$, $T = 0.24$; evolutionary rates in substitutions/site/year of 0.0152 for codon position 1, 0.0142 for codon position 2 and 0.0215 for codon position 3). The evolutionary rates used were estimated from HIV-1 subtype A, B, C, D, and G sequences, collected in the 1990s from Scandinavia (A. Iversen, personal communication). These rates are entirely typical of HIV evolution and are comparable to previously published estimates (e.g. Lemey et al., 2004, 2005). This step produced 3436 alignments of 32 sequences, each 300nt in length. A posterior sample of trees (PST) was then obtained for each alignment using BEAST (Drummond et al., 2002; Drummond and Rambaut, 2003).

In the second step, the simulated sequences and PSTs were used to investigate a hypothetical binary character trait, denoted ‘black’/‘white’. To investigate the null hypothesis, taxa were associated with character traits by randomly sampling without replacement from 16 ‘black’ and 16 ‘white’ states. The assignments of traits to taxa were then applied to the full set of 3436 PSTs, with the assignments being resampled and randomized for each PST. The 3436 PSTs, each with a specified taxon–character association matrix, were analysed in our package BaTS. For each dataset, we calculated the AI, PS and MC statistics for the binary character trait and calculated the p-value of the null hypothesis test. The p-values for each statistic were then collated to generate a cumulative density function (CDF) of p-values for each statistic.

The CDF plots for data sets generated under the null hypothesis are shown in Fig. 2. While the AI plot is a smooth curve that very closely matches the expected unit uniform

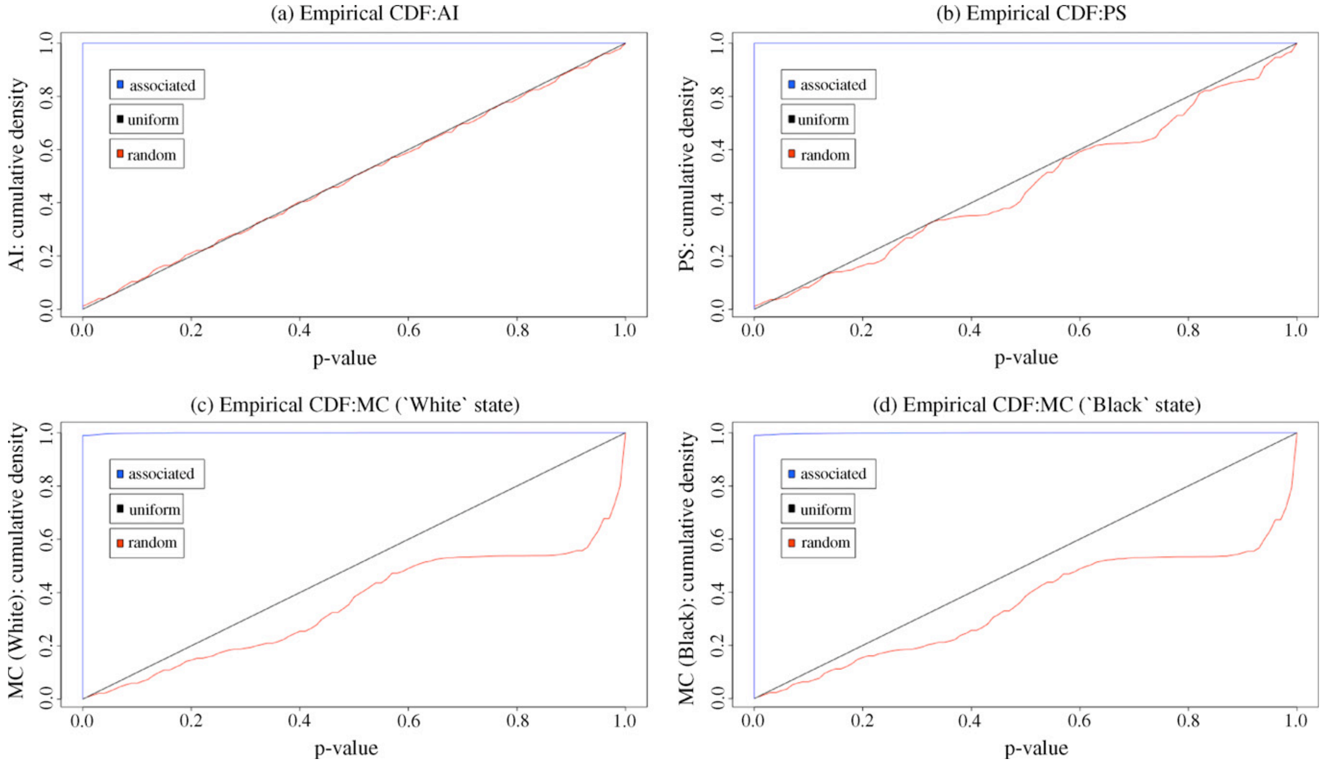


Fig. 2. The expected unit uniform cumulative density function is a black line. The cumulative density functions for each statistic are shown (a) AI statistic, (b) PS statistic, (c) MC(white) statistic, (d) MS(black) statistic. Three thousand four hundred and thirty-six simulated data sets were obtained under two models: the null ‘random’ model of taxon–character association (red) or the ‘completely associated’ model (blue).

distribution, the PS and MC plots deviate from the unit uniform distribution and feature several inflections. Using a two-sample Kolmogorov–Smirnov test (see Lilliefors, 1967), we found that the CDFs of the AI and PS statistics did not depart from the theoretically expected unit uniform distribution; the MC statistic, however, did. The type 1 error rates of the AI and PS statistics, as implemented in a Bayesian MCMC framework, are therefore largely correct.

Finally, to explore the power of each statistic, we repeated the analysis above, but this time specifying a taxon–character matrix corresponding to a very strong phylogeny–trait correlation: the first 16 taxa were assigned the ‘white’ trait and the last 16, ‘black’. Because the null hypothesis should be rejected in every case, the proportion of rejections at the $p = 0.05$ level provides an estimate of the statistical power of our approach on 300nt sequences. The results were as follows: the null hypothesis was rejected for all 3436 simulated data sets when the AI and PS statistics were used. When the MC statistics were used, the null hypothesis was accepted only for 0.35% (white) and 0.5% (black) of the simulated data sets. These results are also contained in Fig. 2, which shows the CDF of *p*-values equals 1.0 for almost all values of *p*. Therefore, all the statistics show high statistical power when a strong tip–trait correlation does exist. The AI and PS statistics show slightly greater statistical power than the MC statistic.

The differences in type 1 error, above, likely result from the fact that AI is a more continuous metric than PS, which in turn can take more possible values than the MC statistic. Here, our simulated phylogenies all had 32 tips, so the range of values that discrete statistics can take on a single tree is limited, hence the MC and PS statistics suffer from a lack of resolution. Furthermore, possible values of the MC statistic are further constrained to the sizes of the monophyletic clades in the tree whose tips all share a trait value. For instance, a perfectly symmetrical tree of 32 tips, 16 of which are ‘white’ and 16 of which are ‘black’, can only take MC(white) and MC(black) values of 1, 2, 4, 8 or 16. Such constraints may explain the significant departure from uniformity by the MC statistic observed here and the lesser degree of departure shown by the PS statistic. Hence, researchers conducting multiple tests with the MC statistic should compare their observed CDF against an appropriate simulated null distribution, rather than the expected uniform distribution.

The statistics investigated here can be considered to span a continuum. At one end, the MC statistic is intuitive, can be scored by hand on a single tree, but has low resolution, reduced power and incorrect type 1 error rates. At the other end, the AI statistic is less intuitive and harder to calculate, but is a better-behaved statistic.

4. Empirical data sets

To evaluate the performance of our method on an empirical set of sequences associated with two trait values, we used two data sets presented by Carrington et al. (2005) that represent the spread of Dengue virus Type 2 (DENV-2) and Dengue virus Type 4 (DENV-4) in the Americas. Each viral sequence is labelled as ‘island’ if it was sampled from a Caribbean island or

Table 1
DENV-2 and DENV-4 dataset results

Statistic	Single ML tree estimate	BaTS estimate (95% HPD CIs)	*p*-value (BaTS null hypothesis test)
DENV-2 data set			
AI	1.33	1.48 (1.07, 1.93)	<0.005
PS	12[a]	11.77 (11,12)	<0.005
MC (island)	4	5.62 (4, 8)	0.185
MC (mainland)	15	16.04 (15, 18)	0.01
DENV-4 data set			
AI	0.397	0.796 (0.336, 1.27)	<0.005
PS	7[a]	8.51 (7, 10)	<0.005
MC (island)	14	16.08 (14, 21)	0.01
MC (mainland)	4	4.66 (4, 7)	0.03

HPD CIs = highest posterior density confidence intervals (credible sets).

[a] As reported in Carrington et al. (2005).

'mainland' if it was sampled from a Central or South American continental nation. In the original paper, the authors calculated the PS statistic on single phylogenies that were estimated using maximum likelihood (ML), and concluded there was sufficient correlation between these geographical characters and the phylogeny to suggest that geography had been a key factor in the spread of Dengue virus in the Americas. We used these published ML trees to also calculate the AI and MC statistics for these data. Next, we reanalysed the DENV-4 and DENV-2 data sets using all three statistics implemented in a Bayesian MCMC framework (see Table 1). Overall, the agreement between the values obtained using the MCMC method and the values calculated from single ML trees was good. Our *p*-values validate Carrington et al. (2005) conclusions, with the exception of the MC(island) statistic for the DENV-2 dataset, which we found not to be significantly larger than that expected by chance.

To evaluate the performance of our method on an empirical dataset with more than two character states, we re-analysed the data set published in Salemi et al. (2005). This study examined HIV-1 sequences isolated from several tissue compartments of the central nervous system immediately after death. Among other aims, the study sought to test the hypothesis of compartmentalization among the seven tissues sampled using the Slatkin–Maddison test (Slatkin and Maddison, 1989). This test was only performed for a subset of the data, so here we have calculated the PS and AI statistics from the ML tree presented in the original paper. Salemi et al. (2005) rejected the null hypothesis of no structure using the Slatkin–Maddison test and also reported the MC size statistic for each tissue sampled. Our results (Table 2) agree very closely with the original analysis; not only in the significance of the PS and MC statistics, but the actual MC sizes also matched closely. However, the AI value obtained from their ML tree fell outside our 95% CIs. It is likely that the large number of polytomies in the ML tree are responsible. Alternatively, due to bias, the ML tree may be a rather poor representative of the PST as a whole.

5. Discussion

Here we have developed a method for investigating phylogeny-tip correlation that accounts for phylogenetic uncertainty. Integrating over the set of all posterior trees is a qualitative and quantitative improvement over single-tree methods. It should produce better estimates of tree statistics and more accurate significance values. While it is reassuring that the published analyses' values fall within our 95% intervals, we also noticed that they did not always fall centrally (Tables 1 and 2). This indicates that statistics estimated from single trees may not accurately reflect the location of the bulk of the posterior probability. Maximum likelihood point estimates are known to be biased in many cases (Edwards, 1972), hence

Table 2
HIV dataset results

Statistic	Single ML tree estimate	*p*-value Salemi et al. (2005)	BaTS estimate (95% HPD CIs)	*p*-value (BaTS null hypothesis test)
AI	0.25[a]	–	1.78 (1.12, 2.51)	0.0056
PS	19[a]	–	19.34 (17, 22)	0.0056
MC (frontal lobe)	15	7 × 10e−7	11.71 (6, 16)	0.01
MC (occipital lobe)	19	3 × 10e−7	18.99 (18, 19)	0.01
MC (meninges)	12	9 × 10e−7	12.32 (12, 13)	0.01
MC (lymph nodes)	10	8 × 10e−7	8.76 (5, 10)	0.0056
MC (temporal lobe)	11	6 × 10e−7	10.98 (10, 11)	0.01
MC (seminal vesicles)	2	0.82	3.19 (2, 5)	0.01
MC (spinal cord)	5	5 × 10e−4	5.01 (5, 6)	0.01

HPD CIs = highest posterior density confidence intervals (credible sets).

[a] Scored from the published ML tree.

the PST may provide a more 'unbiased' estimate of a tree statistic than the value obtained from a single ML tree. In addition, Bayesian MCMC methods are thought to provide a better estimate of phylogenetic accuracy than the bootstrap or jacknife methods commonly used to assess ML trees (Alfaro et al., 2003; Huelsenbeck and Rannala, 2004).

Two situations common in studies of pathogen evolution may be particularly vulnerable to the errors that a single-tree approach can introduce. Firstly, data sets may have weak phylogenetic signal and large phylogenetic error (i.e. viral genetic diversity is low due to very strong negative selection, population bottlenecks, or low mutation rates). In such cases, single tree estimates will have low bootstrap support values and a number of alternative branching orders may be equally plausible. By integrating over the PST, all these possible topologies are taken into account and, importantly, weighted by their posterior probability. Secondly, rapidly-growing or epidemic viral populations are common in viral epidemiology and typically give rise to star-like sample phylogenies. A single tree estimate of such trees may contain numerous polytomies, reflecting the lack of phylogenetic information about branching order near the root. Again, by integrating over the PST we are able to take this uncertainty into account.

Our method is not without limitations. Firstly, the requirement for a PST means that the researcher must first carry out a Bayesian MCMC analysis of the data, which can be time consuming. Secondly, the low resolution of the MC statistic means that some care should be taken in interpreting multiple trials, as previously discussed. Finally, at present, we have only implemented statistics based on tree topology, and we have yet to investigate statistics that use both tree topology and branch length information (e.g. the PD, NTI/NRI and UniFrac measures). However, by placing the PS, AI and MC statistics in a Bayesian inference framework, our method does incorporate statistical variance arising from phylogeny estimation from sequences, which includes variation in both topology and branch lengths. This information is unused by single-tree approaches. It would be useful to consider the PD, NTI/NRI and UniFrac statistics in a similar manner, and we plan to implement and evaluate these metrics in the near future.

Although the two empirical studies presented here focused on spatial characters, the methods could as easily be applied to any other phenotypic traits. Examples might include different risk groups or routes of infection in transmission networks, the different HLA genotypes of infected individuals, or the different disease symptoms of viral infections. The method is readily expandable to larger multi-state data sets, and could also be extended to consider continuously variable traits or ordered discrete states.

Acknowledgements

Thanks to Christine Carrington, Astrid Iversen and Marco Salemi for direct access to their data sets and to Rob Belshaw and Aris Katzourakis for many useful and entertaining discussions and help with code. We are grateful to the two anonymous referees for their helpful comments. This work funded by NERC and The Royal Society.

Full tables of results, including all p-values and computer program code are available on request.

References

Alfaro, M.E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20 (2), 255–266.

Avise, J.C., 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA, 447pp.

Beerli, P., Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. PNAS 98 (8), 4563–4568.

Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., Kadie, C., Carlson, J., Yusim, K., McMahon, B., Gaschen, B., Mallal, S., Mullins, J.I., Nickle, D.C., Herbeck, J., Rousseau, C., Learn, G.H., Miura, T., Brander, C., Walker, B., Korber, B., 2007. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science 315, 1583–1586.

Carrington, C.V.F., Foster, J.E., Pybus, O.G., Bennett, S.N., Holmes, E.C., 2005. Invasion and maintenance of Dengue Virus Type 2 and Type 4 in the Americas. J. Virol. 79 (23), 14680–14687.

Cochrane, A., Searle, B., Hardie, A., Robertson, R., Delahooke, T., Cameron, S., Tedder, R.S., Dusheiko, G.M., de Lamballerie, X., Simmonds, P., 2002. A genetic analysis of Hepatitis C Virus transmission between injection drug users. J. Infect. Dis. 186, 1212–1221.

Drummond, A.J., Rambaut, A., 2003. BEAST v1.0. Available at http://evolve.zoo.ox.ac.uk/beast/.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161, 1307–1320.

Edwards, A.W.F., 1972. Likelihood. Cambridge University Press, Cambridge.

Ewing, G., Nicholls, G., Rodrigo, A., 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. Genetics 168 (4), 2407–2420.

Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Cons. 61, 1–10.

Fitch, W.M., 1971b. Toward defining the course of evolution: minimal change for a specific tree topology. Syst. Zool. 20, 406–416.

Fulcher, J.A., Hwangbo, Y., Zioni, R., Nickle, D., Lin, X., Heath, L., Mullins, J.I., Corey, L., Zhu, T., 2004. Compartmentalization of Human Immunodeficiency Virus Type 1 between blood monocytes and CD4(+) T cells during infection. J. Virol. 78 (15), 7883–7893.

Harvey, P., Pagel, M., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, Oxford, 239 pp.

Holder, M., Lewis, P.O., 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. 4, 275–284.

Holmes, E.C., Zhang, L.Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P., Leigh Brown, A.J., 1995. The molecular epidemiology of HIV-1 in Edinburgh, Scotland. J. Infect. Dis. 171, 45–53.

Holmes, E.C., 2004. The phylogeography of human viruses. Mol. Ecol. 13, 745–756.

Hudson, R.R., Boos, D.D., Kaplan, N.L., 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9 (1), 138–151.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.

Huelsenbeck, J.P., Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53 (6), 904–913.

Jermiin, L.S., Olsen, G., Mengersen, K.L., Easteal, S., 1997. Majority-rule consensus of phylogenetic trees obtained by maximum-likelihood analysis. Mol. Biol. Evol. 14, 1296–1302.

Kemal, K.S., Foley, B., Burger, H., Anastos, K., Minkoff, K., Kitchen, C., Philpott, S.M., Gao, W., Robison, E., Holman, S., Dehner, C., Beck, S., Meyer, W.A., Landay, A., Kovacs, A., Bremer, J., Weiser, B., 2003. HIV-1 in genital tract and plasma of women: compartmentalization of viral sequences, coreceptor usage, and glycosylation. Proc. Natl. Acad. Sci. U.S.A. 100 (22), 12972–12977.

Komatsu, H., Lauer, G., Pybus, O.G., Ouchi, K., Wong, D., Ward, S., Walker, B., Klenerman, P., 2006. Do antiviral CD8+ T cells select hepatitis C virus escape mutants? Analysis in diverse epitopes targeted by human intrahepatic CD8+ T lymphocytes. J. Viral Hepatitis 13, 121–130.

Leigh Brown, A.J., Lobidel, D., Wade, C.M., Rebus, S., Philips, A.N., Brettle, R.P., France, A.J., Leen, C.S., McMenamin, J., McMillan, A., Maw, R.D., Mulcahy, F., Robertson, J.R., Sankar, K.N., Scott, G., Wyld, R., Peutherer, J.F., 1997. The molecular epidemiology of human immunodeficiency virus Type 1 in six cities in Britain and Ireland. Virology 235, 166–177.

Lemey, P., Pybus, O.G., Rambaut, A., Drummond, A.J., Robertson, D.L., Roques, P., Worobey, M., Vandamme, A.M., 2004. The molecular population genetics of HIV-1 group O. Genetics 167, 1059–1068.

Lemey, P., van Dooren, S., Vandamme, A.-M., 2005. Evolutionary dynamics of human retroviruses investigated through full-genome scanning. Mol. Biol. Evol. 22 (4), 942–951.

Lilliefors, H.W., 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J. Am. Statist. Assoc. 62 (318), 399–402.

Lozupone, C., Knight, R., 2005. UniFrac: a new method for comparing microbial communities. Appl. Environ. Microbiol. 71 (12), 8228–8235.

McGrath, K.M., Hoffman, N.G., Resch, W., Nelson, J.A.E., Swanstrom, R., 2001. Using HIV-1 sequence variability to explore virus biology. Virus Biol. 76, 137–160.

Nakano, T., Lu, L., Liu, P., Pybus, O.G., 2004. Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. J. Infect. Dis. 190, 1098–1108.

Pillai, S.K., Kosakovsky Pond, S.L., Lui, Y., Good, B.M., Strain, M.C., Ellis, R.J., Letendre, S., Smith, D., Gunthard, H.F., Grant, I., Marcotte, T.D., McCutchan, J.A., Richmann, D., Wong, K., 2006. Genetic attributes of cerebrospinal fluid-derived HIV-1 *env*. Brain 129, 1872–1883.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13 (3), 235–238.

Rambaut, A., 2001. Phyl-O-Gen. Available at http://evolve.zoo.ox.ac.uk.

Salemi, M., Lamers, S.L., Yu, S., de Oliveira, T., Fitch, W.M., McGrath, M.S., 2005. Phylodynamic analysis of Human Immunodeficiency Virus Type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. J. Virol. 79 (17), 11343–11352.

Sanjuan, R., Codoner, F.M., Moya, A., Elena, S.F., 2004. Natural selection and the organ-specific differentiation of HIV-1v3 hypervariable region. Evolution 58 (6), 1185–1194.

Sheridan, I., Pybus, O.G., Holmes, E.C., Klenerman, P., 2004. High resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. J. Virol. 78, 3447–3454.

Slatkin, M., Maddison, W.P., 1989. A cladistic measure of gene flow measured from the phylogenies of alleles. Genetics 123 (3), 603–613.

Starkman, S.E., MacDonald, D.M., Lewis, J.C.M., Holmes, E.C., Simmonds, P., 2003. Geographic and species association of hepatitis B virus genotypes in non-human primates. Virology 314, 381–393.

Wang, T.H., Donaldson, Y.K., Brettle, R.P., Bell, J.E., Simmonds, P., 2001. Identification of shared populations of human immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. J. Virol. 75 (23), 11686–11699.

Webb, C.O., 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am. Nat. 156 (2), 145–155.

Webb, C.O., Ackerly, D.D., McPeek, M.A., Donoghue, M.J., 2002. Phylogenies and community ecology. Annu. Rev. Ecol. Syst. 33, 475–505.

Wilson, G.A., Rannala, B., 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163 1777–1191.

Wright, S., 1952. The theoretical variance within and among subdivision of a population that is in a steady state. Genetics 37, 312–321.