# Phylogenetic Branch Support Assessment Using Maximum Likelihood Decay Indices and the AU Test

James O. McInerney, Mark Wilkinson, Christopher J. Creevey.

Also: Karen Siu-Ting? Davide Pisani? Omar Rota-Stabelli? Mary O’Connell?

## Abstract

## Introduction

Phylogenetic studies are fundamental to understanding organismal evolutionary history as well as the history of genes, and genomes [REF] and they underpin several areas of biological research, from predicting gene function to informing drug development and forensic analyses. Popular phylogeny inference methods include maximum parsimony (MP), distance matrix (DM), maximum likelihood (ML), and Bayesian inference. These methods are based on explicit mathematical descriptions of how character states transform over time [REF]. The arrival of high-throughput sequencing technologies has revolutionised phylogenetics, leading to an exponential increase in data volume. Research has shifted from single-gene analyses to phylogenomics, where datasets often comprise hundreds or thousands of taxa and millions of nucleotide sites.This genome-scale data initially promised to be a “magic bullet”, seemingly eliminating data availability concerns and stochastic error as limiting factors for resolving deep phylogenetic questions [REF – phylogenomics]. However, this expansion in data quantity has simultaneously amplified challenges related to systematic error and model misspecification, and drastically increased computational complexity. This creates a scenario where more data, while offering unprecedented resolution, also exposes deeper, often unaddressed, methodological flaws and biological complexities [REF]. The initial promise of phylogenomics has been tempered by the reality of increased incongruence and the dangers of systematic biases (biases that amplify with increasing amounts of data(?)). This observation implies that simply accumulating more data is insufficient for achieving accurate phylogenetic inference. Instead, improvements in sophisticated evolutionary models and computationally efficient algorithms are equally, if not more, critical. These advances must be capable of accounting for complex evolutionary processes, such as non-stationary, non-reversible, non-homogeneous conditions, and gene tree-species tree discordance, which become more apparent with larger datasets. The focus therefore shifts from collecting data to intelligently processing, modelling, and critically interpreting its meaning [REF].

Phylogenetic trees are the primary visual representation of inferred evolutionary relationships and are routinely annotated with branch support values [REF]. These values are indispensable for quantifying the uncertainty associated with specific phylogenetic hypotheses and for assessing the reliability of the inferred tree topology. The bootstrap approach of sampling characters in a data matrix, with replacement, introduced by Felsenstein in 1985, remains the most widely used measure of branch support in phylogenetic tree reconstruction [REF]. Despite ongoing debates about its statistical interpretation and known limitations, the original bootstrap method continues to be extensively used. A common empirical guideline, suggested by Hillis and Bull (1993), proposes a 70% bootstrap support threshold as an indication of reliability. However, this threshold was established under very specific, perhaps idealised, conditions, such as equal rates of change and symmetric phylogenies [REF].

A significant limitation of Felsenstein's bootstrap, particularly with the increasing number of taxa (hundreds or thousands) in modern datasets, is its tendency to yield very low support values for deep branches. This issue stems from the method's core methodology: a replicated branch must match a reference branch *exactly* to be counted towards its support value. The presence of “rogue” taxa with unstable phylogenetic positions perhaps because of deep divergence times, or biological factors like convergence, recombination, or even data errors - can severely depress bootstrap support values [REF]. Even if a branch is largely consistent across bootstrap replicates, the misplacement of just one or a few rogue taxa can lead to drastically reduced bootstrap support [REF – Wilkinson?]. The standard approach of pruning these taxa from the analysis is both computationally expensive and statistically questionable [REF]. Furthermore, the bootstrap method is computationally intensive, typically requiring analysis times 100 times longer than phylogenetic inference from the original dataset. While parallelisation and the development of faster algorithms have helped mitigate this, it remains a substantial computational burden for the massive datasets common in phylogenomics.

The statistical meaning of traditional Felsenstein's bootstrap proportions (BPs) is a subject of intense debate, often interpreted as a measure of repeatability rather than a true confidence level [REF]. A significant issue arises with large datasets where FBPs tend to yield very low supports, even for branches that are highly likely to be correct. This suggests that the conventional interpretation and utility of support values, particularly bootstrap percentages, are becoming problematic in the phylogenomic era. This observation necessitates a fundamental re-evaluation of how researchers interpret and communicate phylogenetic confidence. Relying solely on traditional bootstrap values, especially in the context of large datasets, can lead to a significant underestimation of true support or a misinterpretation of underlying phylogenetic signals. The field is compelled to adopt or develop metrics that are more robust to dataset size and better reflect the statistical certainty or the strength of evolutionary evidence for a given branch.

Novel approaches are continually being developed to address the limitations of traditional branch support methods. Transfer Bootstrap Expectation (TBE), for example, is a more recent innovation in phylogenetic bootstrap that measures the presence of inferred branches in replicates using a gradual “transfer” distance, as opposed to the binary presence/absence index of Felsenstein's bootstrap. TBE has been shown to yield higher and more informative support values than the standard approach, while maintaining a very low rate of falsely supported branches. It is particularly robust to high taxon sampling, a scenario where FBP is negatively impacted. The Bayesian Bootstrap offers a Bayesian formulation of the phylogenetic bootstrap where sites are assigned uninformative prior probabilities. This approach allows branch support to be interpreted directly as a posterior probability, which can be more intuitive. Simulations and analyses of low-homoplasy viral and nonviral datasets show that Bayesian bootstrap support is easier to interpret, providing high support for branches that are very likely to be correct, and yielding higher expected support for branches with single mutations (~90% compared to ~63% for FBP). Approximate Likelihood Ratio Tests (aLRT) provide faster alternatives to traditional bootstrap for assessing local branch support. While more computationally efficient, their approximate nature means they are not universally accepted by all researchers. Another method proposed as an alternative to Felsenstein's bootstrap is aBayes support, though it is noted to be difficult to apply to large datasets and sensitive to model misspecification [Ref - Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes].

The Approximately Unbiased (AU) test [REF] is a statistical test developed to reduce bias in general hypothesis testing, particularly in the context of maximum-likelihood tree selection. It employs a newly devised multiscale bootstrap technique. The AU test aims to provide greater accuracy, and a simpler implementation compared to earlier methods like the Shimodaira-Hasegawa (SH) [REF] test. A key feature of the AU test is its ability to adjust for selection bias, which commonly arises when comparing multiple trees simultaneously and can lead to overconfidence in incorrect tree topologies.26 The test establishes a null hypothesis that the distributions of site likelihoods for different candidate trees (or roots) are statistically equivalent, and it returns a p-value for each tested tree or root.

The primary application of the AU test is when a user obtains a set of trees, the AU test identifies those phylogenetic hypotheses that are not significantly worse reconstructions of the phylogeny than the maximum likelihood tree and those that are. While the AU test is designed to reduce bias, its asymptotic theoretical foundation assumes a smooth boundary for hypothesis testing, which may not perfectly hold for the complex polyhedral convex cone formed by tree selection problems [REF An Approximately Unbiased Test of Phylogenetic Tree Selection. Shimodaira]. The impact of dataset size on the AU test's performance is implicit, as systematic bias, which the AU test aims to mitigate, has an increasing effect on inferences from larger phylogenetic datasets. Empirical studies show that instability in tree topology, even if not always statistically significant by the AU test, occurs frequently when new taxa are added to large datasets [need REF for this].

A recurring observation across the discussion of bootstrap and AU tests is a fundamental trade-off. Bootstrap is broadly applicable and conceptually simple but computationally demanding and its statistical interpretation is ambiguous. The AU test aims for reduced bias and higher accuracy by employing complex multiscale bootstrap techniques, which also involve significant computational resources. This implies that there is no single “optimal” branch support method universally applicable to all phylogenetic analyses. It is necessary to carefully navigate this trade-off, making informed decisions based on their specific dataset characteristics (size, complexity), available computational resources, and the precise biological questions being addressed.

The AU test was specifically developed to address “selection bias” that can lead to “overconfidence in the wrong trees” when numerous tree topologies are compared [REF]. This concern is echoed in the context of large phylogenomic datasets, where systematic errors can inflate support values for incorrect topologies, meaning that even high support values may not reflect true accuracy. This suggests that high support values, particularly those approaching 100% in large datasets, are not always reliable indicators of phylogenetic truth. Instead, they can be artifacts of methodological biases, model misspecification, or the sheer volume of data reducing stochastic error without addressing systematic error [REF]. The means the emphasis when analysing large datasets shifts from merely *obtaining* high support to *critically interpreting* its meaning within the broader context of the data and the analytical methods employed.

The Decay Index, widely known as Bremer support (Bremer, 1988) or sometimes referred to as patristic difference (PD) support metric [REF-Kluge], is a method for quantifying branch support in parsimony-based phylogenetic analyses. It measures the robustness of a clade by calculating the difference in tree length (i.e., the number of evolutionary steps or mutations) between the most parsimonious tree (MPT) that *lacks* that specific clade and the overall MPT. Essentially, it quantifies how many additional evolutionary steps are required to collapse a particular branch in the tree, or to support an alternative topology that contradicts that branch. A higher decay index value indicates stronger support for the branch, as it implies a greater cost (more steps) to remove or contradict it.

The program TreeRot [REF] is designed to assist in the determination of decay indices by generating specialised command files for the phylogenetic analysis program PAUP\* [REF]. This process involves constructing “constraint statements” for each internal node (or clan [REF]) in a given reference tree (*e.g.,* the MPT). Subsequently, the software searches for the shortest trees that are *inconsistent* with these specified constraints (*i.e*., trees that do not contain the clan in question). The decay index for a node is then calculated as the difference in tree length between the shortest inconsistent tree(s) and the shortest unconstrained tree. This constraint-based approach is significantly more efficient and effective for determining higher decay indices compared to simply finding all trees of incrementally longer lengths and then examining their strict consensus [REF].

Lee and Hugall proposed a Bremer support-like approach for maximum likelihood [REF]. However, their approach related to partitioned datasets. In the more straight-forward context, likelihood support for a particular clade is quantified as the difference in log-likelihood scores between the optimal tree topology that *includes* that clade and the optimal topology that *excludes* it. This provides a likelihood-based equivalent to the parsimony “steps” difference. Therefore, a theoretically sound method exists, but its detailed, standardised implementation in widely accessible software is lacking. This practical barrier could hinder the widespread adoption of a potentially valuable support metric, despite its theoretical appeal in a maximum likelihood framework.

## MLDecay

To address the practical gap between theoretical frameworks and accessible implementations in phylogenetic support assessment, we have developed MLDecay, a Python command-line tool that implements maximum likelihood-based decay indices for phylogenetic branch support assessment. The software extends the classical parsimony-based Bremer support concept to the ML framework by automating the calculation of likelihood differences between optimal ML trees and constrained alternatives where specific clades are forced to be non-monophyletic. MLDecay is available under an MIT Open Access licence from <https://github.com/mol-evol/MLDecay/> .

MLDecay leverages PAUP\* as its phylogenetic engine and automates the ML decay calculation process through five key steps: (1) finding the optimal ML tree and its likelihood score, (2) for each internal branch in the ML tree, defining a constraint that forces the taxa in that clade to not form a monophyletic group using PAUP\*'s converse=yes constraint, (3) searching for the best ML tree under this reverse-constraint and recording its likelihood, (4) calculating the difference in log-likelihood between the unconstrained ML tree and each constrained tree, and (5) performing an Approximately Unbiased (AU) test to statistically compare the unconstrained ML tree against all the constrained alternative trees.

The software extends beyond traditional decay index calculations by incorporating several advanced features designed for modern phylogenomic analyses. MLDecay supports DNA, protein, and binary discrete morphological data with flexible model specification (GTR, HKY, JTT, WAG, Mk models) and options for gamma-distributed rate heterogeneity (+G) and proportion of invariable sites (+I). It allows fine-grained control over model parameters and provides an option for users to supply custom PAUP\* blocks for complex model or search strategy definitions. Importantly, MLDecay addresses the need for complementary support measures by optionally performing bootstrap analysis alongside ML decay calculations, enabling direct comparison between these fundamentally different approaches to branch support assessment.

MLDecay also has a clade-specific site-specific likelihood analysis capability. When the --site-analysis option is enabled, the software identifies which alignment positions support or conflict with each branch in the tree by calculating site-by-site likelihood differences between the ML tree and constrained alternatives. This analysis generates comprehensive output including supporting vs. conflicting site counts, support ratios, and detailed visualizations showing the distribution of site-specific evidence. This functionality directly addresses the challenge of understanding the underlying phylogenetic signal contributing to branch support, moving beyond simple summary statistics to examine the character-level evidence that may reveal alignment errors, recombination events, or differing selective pressures.

MLDecay is particularly well-suited for phylogenomic datasets where traditional bootstrap support may be inadequate or misleading. ML decay indices provide an alternative perspective on branch support that may be less susceptible to certain systematic biases than bootstrap methods, particularly in phylogenomic contexts where systematic biases can inflate confidence values.

The primary outputs of MLDecay include likelihood differences (ML decay values), AU test p-values, and optionally bootstrap support values for each internal branch. The log-likelihood difference is calculated as the constrained tree likelihood minus the ML tree likelihood, where larger positive values indicate stronger support for the original clade. The AU p-value tests the null hypothesis that the ML tree is not significantly better than the constrained alternative; low p-values (e.g., < 0.05) lead to rejecting this null hypothesis, providing statistical support for the original clade's monophyly. The software generates multiple annotated tree files for visualization, including trees with AU p-values, likelihood differences, and combined annotations suitable for programs like FigTree [REF].

The interpretation of MLDecay results requires careful consideration of both the magnitude of likelihood differences and their statistical significance as assessed by the AU test. Clades with large positive likelihood differences and low AU p-values should be considered well-supported, but these values should be interpreted within the broader context of model adequacy and potential systematic errors. The integration of site-specific analysis results provides additional layers of evidence, allowing researchers to assess whether support is broadly distributed across the alignment or concentrated in particular regions that may warrant further investigation.

The software requires Python 3.8+ with standard scientific libraries (BioPython, NumPy) and a working PAUP\* installation. It is compatible with Unix-like operating systems (Linux, macOS) and is designed for phylogeneticists working with large datasets who need alternative perspectives on branch support that complement bootstrap methods, or when understanding the distribution of phylogenetic signal across alignments is critical for interpreting evolutionary relationships.

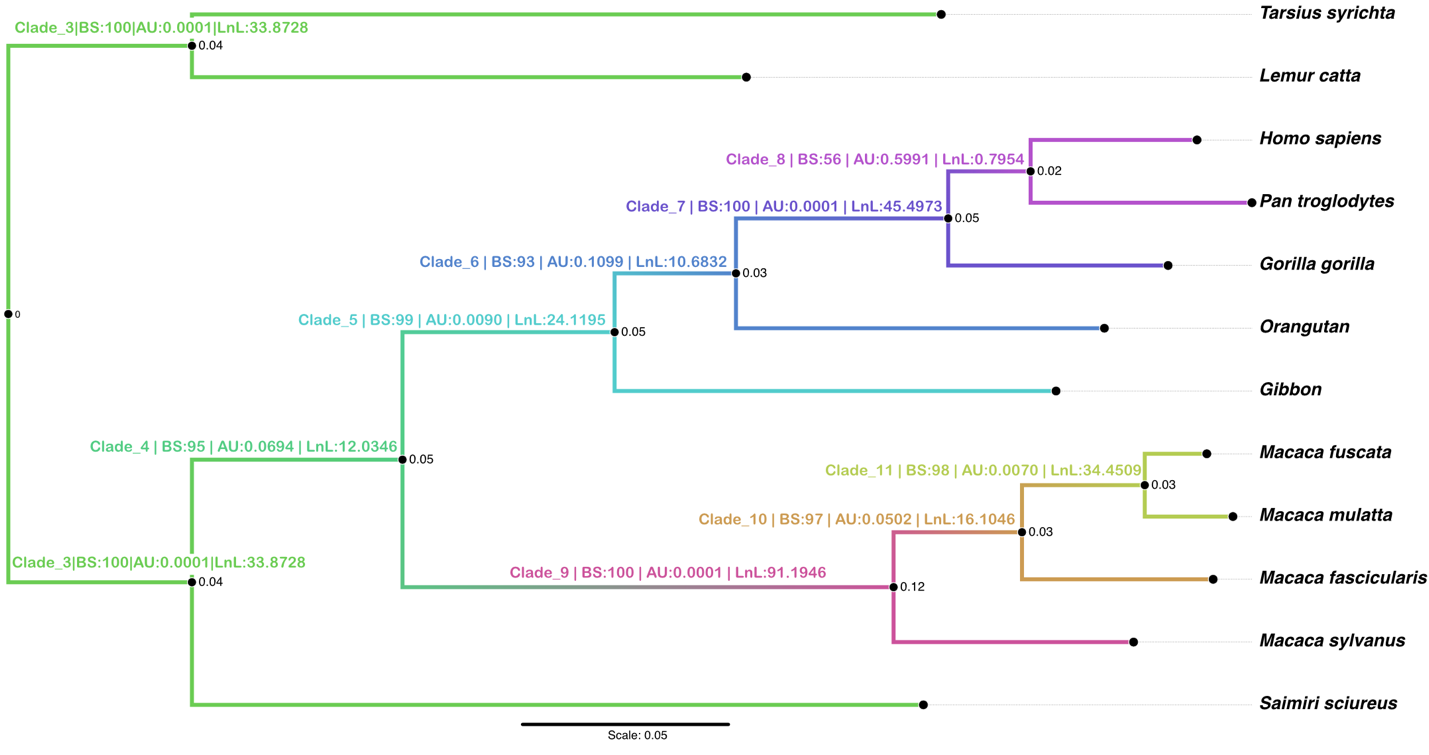
## Exemplar Analysis Results

**Figure 1** presents the results of a basic analysis of primate mitochondrial DNA phylogeny [REF] where we can explore of the relationship between Bootstrap Support, Δlog-likelihood, branch length and AU-Decay probability (???). This analysis was not designed to identify the best-fitting model of nucleotide substitution; instead, we used the Jukes-Cantor (1969) model simply to illustrate how various support measures interact. Internal branch lengths are shown on the tree, rounded to two decimal places.

Starting with the clade that groups humans and chimpanzees, the internal branch length is short - just 0.2 substitutions per site. This clade has a Δlog-likelihood score of 0.7954, receives weak bootstrap support (56%), and is not supported by the approximately unbiased (AU) test (P = 0.59). In contrast, the next internal branch - uniting humans, chimpanzees, and gorillas - tells a different story. The branch length is more than twice as long, and the Δlog-likelihood score is much higher at 45.4. This clade receives 100% bootstrap support, and the AU test strongly supports it (P ≈ 0). The clade including the orangutan has a shorter internal branch (0.03 substitutions per site) and a Δlog-likelihood score not much higher than that for the human–chimpanzee clade. Despite receiving 93% bootstrap support, it is not supported by the AU test (P = 0.1099). This illustrates an important case where bootstrap support appears strong, yet the AU test does not confirm it. Clade 5, which includes the gibbon, shows both high bootstrap support (99%) and AU test support (P = 0.009).

Clade 4 also has strong bootstrap support (95%) but fails the AU test (P = 0.069). Another interesting case is Clade 10, which receives 97% bootstrap support but narrowly fails the AU test. Its internal branch length is 0.03, and the Δlog-likelihood score is 16.1. Comparing Clades 4 and 10 is instructive. Clade 4 has a longer internal branch (0.05) than Clade 10 (0.03), yet its Δlog-likelihood score is lower.

These examples highlight that while bootstrap support, Δlog-likelihood (or MLDecay) scores, branch lengths, and AU test results are all interconnected, no single measure gives a complete picture. Instead, a full understanding emerges only when all metrics are considered together.



**Figure 1**: Primate phylogeny inferred using Maximum Likelihood, with internal branch lengths indicated on nodes, and MLDecay clade numbers, Bootstrap Support values, AU test result and log-likelihood difference all indicated on internal branches.

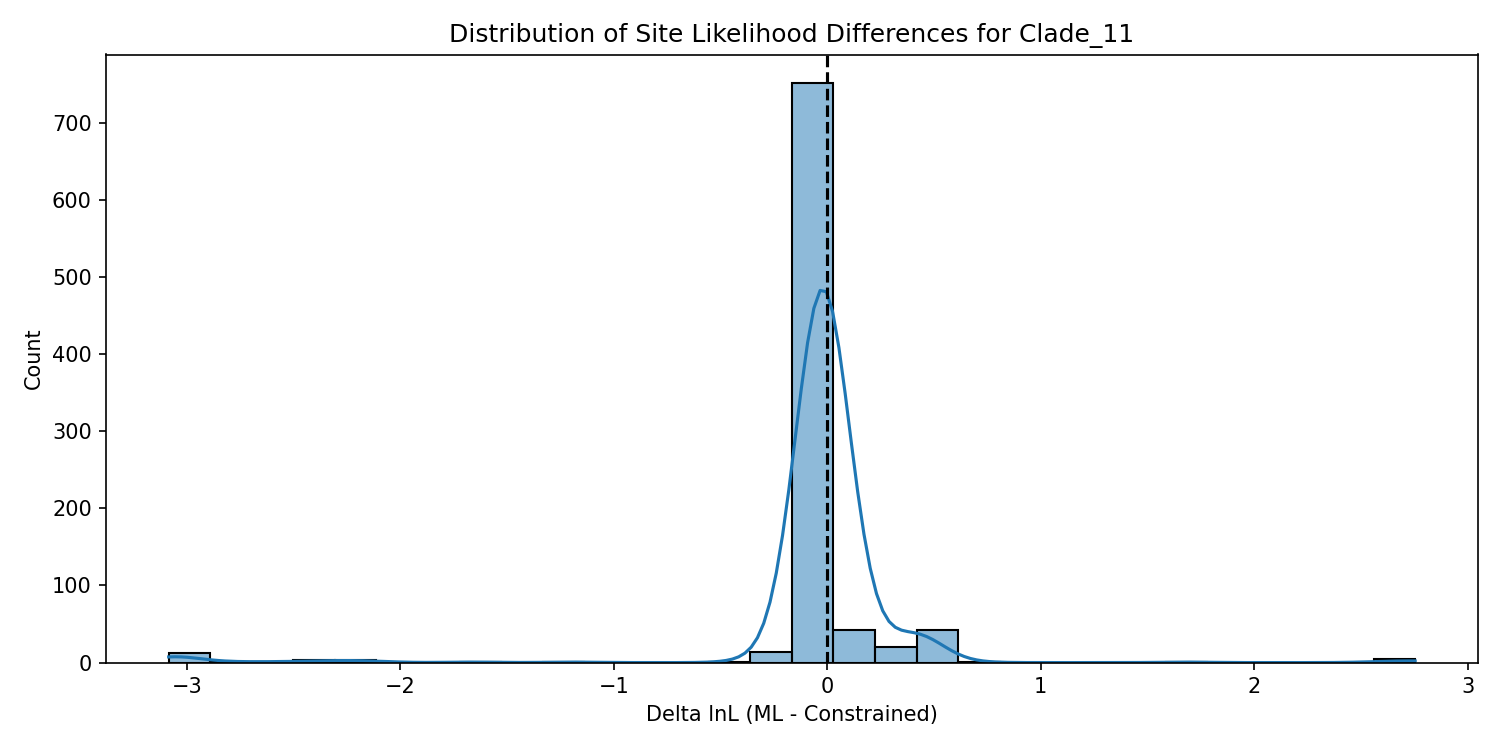
The plot in figure 2 presents the results of a site-specific likelihood difference analysis for a particular clade in the phylogenetic tree - clade 11, which groups *Macaca fuscata* with *M. mulatta*. Each bar in the figure represents the difference in log-likelihood for a given site when comparing the optimal (unconstrained) tree to a tree in which this clade is constrained. Green bars indicate that a site supports the unconstrained tree more strongly; the bar’s length reflects how much better the site fits the optimal tree. Red bars show sites that favour the constrained tree, meaning their likelihood scores are higher on the best tree where clade 11 is forced to be absent. This type of analysis could be useful for detecting recombination or horizontal gene transfer. For instance, a cluster of consecutive red bars may suggest a region of the alignment that evolved under a different tree topology, consistent with such events. In addition, the MLDecay program will report the relative supports for this clade by reporting the number of sites favouring the unconstrained clade (752 in this analysis), the number that support the best tree that doesn’t contain the clade (146 in this case), the raw ratio of support (5.15) and a ratio of support and conflict that is weighted by likelihood scores (1.87 in this case).

The MLDecay program will output a site-specific figure for every clade on the tree and these plots can be compared to understand whether support comes from different parts of the matrix for different groups on the tree.



**Figure 2**: Site-specific, clade-specific analysis of support and conflict for a focal clade. This plot refers to clade 11 on Figure 1. The green bars indicate nucleotide columns where the likelihood is higher for the unconstrained (maximum likelihood) tree, whereas the red bars indicate nucleotide columns that have a higher likelihood for the best tree that doesn’t contain clade 11.

The distribution shown in Figure 3, which is derived from Figure 2, indicates that most sites in the data matrix with a likelihood score difference between the constrained and unconstrained trees showed only a small difference. While most sites offered marginal support, a few outliers strongly favoured one tree over the other. As expected, the maximum likelihood (unconstrained) tree received broader overall support. Of the 752 sites that favoured it at this internal node, more than 725 showed only a modest preference - less than 1 log-likelihood unit. However, 17 sites provided strong support, with differences exceeding 3 log-likelihood units in favour of the unconstrained tree. In contrast, the constrained tree was supported by fewer sites - just 146 in total. Although these sites tended to have higher average Δlog-likelihood scores compared to most sites supporting the optimal tree, none exceeded a difference of 3 log-likelihood units. This pattern highlights that strong overall support for the optimal tree is largely due to the cumulative effect of many small contributions, rather than a few highly influential sites. The MLDecay program will produce an output of this kind for every internal branch of the optimal tree.



**Figure 3**: Distribution of site likelihood differences for Clade 11. The differences in likelihood score are binned into an appropriate number of bins and the counts are shown on the left.

## Discussion

The phylogenomic era has fundamentally challenged traditional approaches to assessing branch support in phylogenetic inference [REF]. While the accumulation of vast genomic datasets initially promised to resolve deep evolutionary questions definitively, the reality has been more complex [REF]. Large datasets can amplify systematic biases and model misspecification, leading to inflated support values that may not reflect true phylogenetic accuracy. This phenomenon necessitates a critical evaluation of how we measure and interpret branch confidence in modern phylogenetic analyses.

MLDecay addresses a gap between theoretical frameworks and practical implementation in phylogenetic support assessment. While Lee and Hugall (2003) established the theoretical foundation for maximum likelihood-based decay indices, no readily accessible software has implemented this approach until now. Our tool provides researchers with a straightforward method to calculate ML decay indices using reverse constraints, offering an alternative perspective on branch support that complements traditional bootstrap, jackknife or Bayesian Inference approaches.

The fundamental strength of MLDecay lies in its ability to quantify the likelihood cost of rejecting specific phylogenetic hypotheses. Unlike bootstrap support, which can be severely depressed by rogue taxa placement regardless of overall clade stability, ML decay indices measure the direct statistical evidence for each branch. This approach is particularly valuable for phylogenomic datasets where traditional bootstrap values may be misleadingly low despite strong underlying signal. Our exemplar analysis demonstrates that bootstrap support and ML decay values can provide conflicting assessments of the same branches, highlighting the importance of considering multiple support metrics.

The site-specific analysis capability of MLDecay can also provide valuable information to assist in understanding the sources of phylogenetic support. By identifying which alignment positions contribute to, or conflict with, each branch, it may be possible to detect potential recombination events, horizontal gene transfer, or alignment errors that may compromise phylogenetic inference. This granular view of support distribution moves beyond summary statistics to highlight character-level evidence underlying tree topology, enabling more informed interpretation of phylogenetic results.

However, it is important to emphasise that no single support measure provides a complete picture of phylogenetic confidence. The relationship between branch length, likelihood differences, bootstrap support, and AU test results is complex and likely to be context dependent. Some groupings on a tree might have a relatively straightforward relationship between branch length, bootstrap support, DLog-Likelihood and AU test, most likely when there are many supporting sites and few conflicting sites. However, we expect that when support at any node is complicated with a plurality of supporting and conflicting signals, then it is important to tease these apart.

Our primate analysis illustrates cases where high bootstrap support occasionally coincides with non-significant AU test results, demonstrating that different metrics can capture distinct aspects of phylogenetic uncertainty. This observation reinforces the need for pluralistic approaches to branch support assessment. Modern phylogenetic analysis requires a nuanced understanding of the interplay between data characteristics, model adequacy, and support metrics [REF]. It is important to move beyond reflexive reliance on any single measure and instead adopt integrated approaches that consider multiple lines of evidence. MLDecay is designed to contribute to this pluralistic framework by providing an accessible implementation of likelihood-based decay indices, enabling researchers to triangulate branch confidence using complementary methodologies.

The challenges facing phylogenomics include systematic bias amplification, model misspecification, and the interpretation of high support values in large dataset. These problems have not been solved so far by any single methodological advance. Instead, progress on these issues likely requires the development and integration of diverse analytical tools that capture different aspects of phylogenetic support. MLDecay represents one component of this broader toolkit, designed to enhance rather than replace existing approaches to assess confidence in groups on trees.

References