Detecting the Anomaly Zone in Species Trees and Evidence for a Misleading Signal in Higher-Level Skink Phylogeny (Squamata: Scincidae).

Charles W. Linkem, Vladimir N. Minin, and Adam D. Leaché

Abstract—The anomaly zone, defined by the presence of gene tree topologies that are more probable than the true species tree, presents a major challenge to the accurate resolution of many parts of the Tree of Life. This discrepancy can result from consecutive rapid speciation events in the species tree. Similar to the problem of long-branch attraction, including more data via loci concatenation will only reinforce the support for the incorrect species tree. Empirical phylogenetic studies often employ consistent-based species tree methods to avoid the anomaly zone, but to this point these studies have not had a method for providing any direct evidence that the species tree is actually in the anomaly zone. In this study, we use 16 species of lizards in the family Scincidae to investigate whether nodes that are difficult to resolve place the species tree within the anomaly zone. We analyze new phylogenomic data (429 loci), using both concatenation and coalescent-based species tree estimation, to locate conflicting topological signal. We then use the unifying principle of the anomaly zone, together with estimates of ancestral population sizes and species persistence times, to determine whether the observed phylogenetic conflicts is a result of the anomaly zone. We identify at least three regions of the Scincidae phylogeny that provide demographic signatures consistent with the anomaly zone, and this new information helps reconcile the phylogenetic conflict in previously published studies on these lizards. The anomaly zone presents a real problem in phylogenetics, and our new framework for identifying anomalous relationships will help empiricists leverage their resources appropriately for investigating and overcoming this challenge. [Anomalous gene trees; IDBA; incomplete lineage sorting; phylogenetics; probes; sequence capture; ultraconserved elements; UCE]
FIGURE 1. The lengths of branches X and Y in coalescent units in the species tree determine the probability of the gene tree topology. For branches under the anomaly zone curve, the symmetric AGT will have a higher probability than the asymmetric gene tree that matches the species tree.

three possible AGTs. With the addition of a single taxon, the five-taxa species tree has multiple sets of x and y and z internode branches that can exist as many as 45 AGTs (Rosenberg and Tao 2008). The calculation of the multidimensional anomaly zone in trees larger than five taxa is impractical, but a conservative simplification of the theory can be used for any species tree (Rosenberg 2013).

Species tree topologies of five taxa or more can be broken up into sets of four-taxon trees, which can individually be used in the anomaly zone calculation of equation (1). Rosenberg (2013) showed that focusing on sets of consecutive internal branches consistent with a four-taxon topology is a conservative estimate of the presence of the anomaly zone in any species tree. If the set of internodes fits the anomaly zone for the four-taxon case, at least one AGT exists, though more AGTs may occur due to nearby branches not considered in the isolated calculation (Rosenberg 2013). The unifying principle of the anomaly zone, that the anomaly zone condition can be checked for any four-taxon case within a larger phylogeny independently of other subtrees, allows for the estimation of the anomaly zone in trees of any size.

The theoretical predictions of the anomaly zone are well characterized (Rosenberg and Tao 2008; Degnan and Rosenberg 2009; Rosenberg and Degnan 2010; Degnan et al. 2012a,b; Degnan 2013), and simulations have identified situations where certain phylogenetic methods succeed or fail under the anomaly zone (Steel and Rodrigo 2008; Huang and Knowles 2009; Liu and Edwards 2009; Liu et al. 2010b; Liu and Yu 2011; Liu et al. 2015). Empirical studies of genomic data have shown that incomplete lineage sorting can be pervasive throughout the genome in rapidly speciating groups (Suh et al. 2015). However, an empirical example of the anomaly zone has yet to be demonstrated. The lack of empirical evidence for the anomaly zone has led to doubt about the influence of the anomaly zone on real data (Huang and Knowles 2009; Tonini et al. 2015), and the utility of coalescent methods for phylogenetic inference has been questioned (Gatesy and Springer 2014; Springer and Gatesy 2014). An investigation of the anomaly zone in an empirical setting requires an accurate species tree topology and estimates for ancestral branch lengths and population sizes, parameters that can be inferred accurately with hundreds of loci. Until recently, obtaining loci at this magnitude was not feasible for most nonmodel organisms, but new methods for obtaining large phylogenomic data sets are quickly changing the scale and scope of phylogenetic studies (Faircloth et al. 2012; McCormack et al. 2012; Song et al. 2012).

Here we present an empirical example of the anomaly zone in a diverse radiation of lizards, the Scincidae. Using theoretical expectations, we define the set of species tree branches expected to generate AGT based on the multispecies coalescent, and we apply these predictions to an empirical species tree. We use a new phylogenomic data set collected using sequence capture of ultraconserved elements (Faircloth et al. 2012) and protein-coding genes (Wiens et al. 2012) to estimate the species tree, branch lengths, and population sizes required for identifying the anomaly zone in an empirical phylogeny.

Scincidae

Scincidae is the largest family of lizards comprising >25% of all lizard species. This globally distributed family of 1583 species (Uetz 2015) can be found from deserts to rain forests, across continents and on tiny Pacific islands (Greer 1970). This relatively ancient group (80–110 Ma; Mulcahy et al. (2012)) exhibits extreme variation in body-form evolution with repeated limb loss (Greer 1991; Skinner et al. 2008) and variation in reproductive biology through multiple transitions to live birth (Blackburn 1982). The first major division of higher-level skink taxonomy was largely based on skull morphology, dividing the family into four subfamilies: Acontinae, Feylininae, Lygosominae, and Scincinae (Greer 1970). Synapomorphies were defined for all...
but Scincinae, which was described as possessing the “primitive” form from which the other forms evolved, indicating Scincinae was not a natural group (Greer 1970). Molecular phylogenies have largely supported the morphological hypotheses of Lygosominae and Acantinae monophyly but have found Lygosominae and Feylininae nested within Scincinae (Whiting et al. 2003; Brandley et al. 2005, 2012; Wiens et al. 2012; Pyron et al. 2013; Lambert et al. 2015). A few molecular studies found paralogy in Lygosominae (Whiting et al. 2003; Siler et al. 2011), though the majority of molecular phylogenies (Wiens et al. 2012; Pyron et al. 2013; Reeder 2003; Lambert et al. 2015) and a series of morphological synapomorphies (Greer 1970, 1986) support the monophyly of this subfamily. Feylininae is no longer considered a valid subfamily and the species within it are part of the broader paraphyletic Scincinae (Whiting et al. 2003; Brandley et al. 2005). The relationships among the major groups within Scincinae have remained difficult to decipher (Brandley et al. 2005, 2012; Wiens et al. 2012; Pyron et al. 2013). Most studies have found Scincinae to be paraphyletic with the genus Brachymeles sister to Lygosominae, but two recent papers (Pyron et al. 2013; Lambert et al. 2015) found a monophyletic Scincinae.

Molecular phylogenies of skinks have sampled up to 44 genes for a few taxa (Wiens et al. 2012; Lambert et al. 2015) or a few genes for up to 663 taxa (Pyron et al. 2013) but still lack resolution or present strong conflict for many relationships. The deep branches across this group are often found to be short when estimated from individual genes and concatenated genes, indicating diversification may have been rapid in this group. Strong support for different relationships from concatenated gene trees and coalescent species trees (Lambert et al. 2015) in combination with a history of rapid diversification indicates that speciation history may fit the demographic parameters consistent with the anomaly zone.

**Materials and Methods**

Identification of the causes of gene tree–species tree conflict in empirical phylogenies requires a large number of sampled genes on taxa that span difficult nodes (Liu and Edwards 2009). To accomplish this, we use a sequence-capture next-generation sequencing approach to obtain 429 independent loci for all taxa of interest.

**Taxon Sampling**

Sampling was focused on 15 species of skinks that span the deep nodes in the Scincidae tree where topological conflict is high. From Scincinae we include Brachymeles bonitae, Chalcides ocellatus, Eurylepis taminolus, Mesoscincus mangua, Ophiomorus radilis, Plestiodon fasciatus, and Scincus sicnus, which represent the broad diversity in this difficult-to-resolve subfamily. These species have been used in prior studies of skink relationships based on Sanger sequencing (Brandley et al. 2005, 2012; Pyron et al. 2013) and have shown very short internode lengths and multiple alternative topologies suggesting AGTs may be present. A single Acontinae sample, Typhlosaurus sp., is used to represent this well-supported subfamily (Lamb et al. 2010). We include at least one sample of four of the five groups in Lygosominae (missing Egeria group); Malbaya unimarginata for the Malbaya group; Lygosoma brevicaudus for the Lygosoma group; Eunoa carradavensis for the Eunygia group; and Lobelia elegans, Sphenomorphus trigeditus, Sphenomorphus variegatus, and Tyloloscincus parvus for the Sphenomorphus group. Previous studies (Honda et al. 2003; Reeder 2003; Skinner 2007; Skinner et al. 2011) have shown variation in the relationships between the Lygosoma, Malbaya, and Eunygia groups, some with short internode lengths. A single out-group taxon, Xantusia vigilis, was chosen to root phylogenetic analyses.

**Probe Design, Library Preparation, Sequencing, and Data Processing**

We designed a unique set of sequence-capture probes targeting a combination of Ultraconserved Element (UCE) loci and loci used in squamate phylogenetics. The subset of UCE probes used is derived from the Tetrapsida-UCE-Str1 probe set (Faircloth et al. 2012) that most closely matches the Anolis genome (Alfoldi et al. 2011) and the Sceloporus occidentalis genome (Genomic Resources Development Consortium et al. 2014). Additionally, we designed 120-bp probes with 2× tiling (60 bp overlap) for the 44 genes used in the Squamate Tree of Life project (Wiens et al. 2012). This resulted in a sequence-capture probe set targeting 585 single copy loci that can be used across squamates. Details of probe design and synthesis are available in the Supplementary Material, available on Dryad at http://dx.doi.org/10.5061/dryad.sfs69.

Whole genomic DNA was broken through sonication using a Bioruptor Pico (Diagenode Inc.) and quality was assessed using an Agilent TapeStation 2200 (Agilent Tech.). Samples were sequence at the QB3 facility at UC Berkeley on an Illumina HiSeq2500 on a single-lane, 150-bp, paired-end rapid run. Details of library preparation, sequence-capture hybridization, and sequencing are available in the Supplementary Material.

Sequence data were processed using Illumiprocessor v.2.0 (Faircloth 2013) and Trimomatic (Bolger et al. 2014) to remove low-quality reads, trim low-quality ends, and remove adapter sequence. De novo assembly for each species was conducted with the iterative de Brujin graph short-read assembler (IDBA) (Peng et al. 2010) over a set of k-mer
values from 50 to 90 with a step length of 10. We used Phylocoe (Faircloth et al. 2012; Faircloth 2014) to align species-specific contigs to the set of probes and create an SQL database of config-to-probe matches for each taxon. We queried this database to create the data sets for the 429 loci that were sequenced across all taxa. Expanded details of data processing are available in the supplementary Materials. Data sets and analysis results are available on Dryad. Sequences are available on GenBank (KU343288–KU350151).

Model Testing, Gene Trees, and Concatenation Analyses
We identified the models of substitution in the 95% credible set of the Bayesian information criterion (BIC) for each locus using jModelTest v2.1.5 (Guindon and Gascuel 2003; Darriba et al. 2012). The model with the lowest BIC score was chosen as the preferred model (supplementary Material). Each locus was evaluated for the number of parsimony-informative sites, number of constant sites, and number of variable sites using PAUP* v. 4b10 Swofford 2003; Supplementary Material). A maximum-likelihood phylogenetic analysis was conducted on each locus using RAxML v7.2.8 (Stamatakis 2006) with 1000 rapid-bootstrap replicates with the GTRGAMMA model.

All loci were concatenated into a single alignment for maximum likelihood (ML) (RAxML) and Bayesian analysis (ExaBayes). The concatenated data set was partitioned by locus (429 partitions) for both analyses. The ML analysis was conducted with the GTRGAMMA model with 1000 rapid-bootstrap replicates. ExaBayes analyses were run with the GTRGAMMA model with branch lengths linked across partitions and a parsimony starting tree with heated chains using different starting trees than the cold chain. Four independent runs were conducted, each with four chains, sampling every 500 generations. ExaBayes runs continued until the termination condition of mean topological difference less than 5% with at least 500,000 generations was met. Posterior distributions of trees were summarized with the consensus script, and posterior sample of parameters was assessed with Tracer v1.5 (Rambaut and Drummond 2007) and combined with the postProcParam script.

Species Tree Estimation
Due to the large number of genes sampled, we limit our species tree estimation to a summary statistic approach. Species tree accuracy in summary statistic approaches is dependent on gene tree accuracy (Huang and Knowles 2009; Mirarab et al. 2014) since the methods rely solely on the structure of the fully resolved gene trees. Loci with few informative sites, often seen in next generation sequencing data sets, may not give strong support for all splits in the gene trees. This can potentially bias the species tree estimate and analyses relying on the species tree topology, such as the identification of the anomaly zone. We use the maximum pseudo-likelihood estimation of species trees MP-EST v1.4 because it can accurately estimate the species tree topology despite the anomaly zone and in the presence of gene tree estimation error (Liu et al. 2010a). We accounted for gene tree and species tree uncertainty by running MP-EST using each iteration of the 1000 maximum-likelihood bootstrap replicates from RAxML. We created 1000 new tree files, consisting of 429 trees, one bootstrap replicate of each locus. MP-EST was run on each of these files, and an extended majority-rule consensus (eMRC) tree of the resulting species trees was calculated using sumtrees in Dendropy v4.0.3 (Sukumaran and Holder 2010). The support values from this approach are a summary of bootstrap replicates from the gene trees to account for error in gene tree estimates but should not be considered a true bootstrap replicate of the species tree.

Identifying Anomalous Nodes
We use the unifying principle of the anomaly zone (Rosenberg 2013) to determine which, if any, parts of the Scincidae eMRC species tree topology should produce AGTs. This procedure requires an estimate of internal branch lengths in coalescent units. We estimated branch lengths using two different methods. First, we used the branch lengths estimated by MP-EST, which jointly estimates internal branch lengths in coalescent units (based on λ = 2λ/θ) while maximizing the pseudo-likelihood of the species topology given the set of triplet topologies for each gene tree (Liu et al. 2010a, Equations 6–8). Internal branches estimated by MP-EST may be shorter than expected when gene tree error is high (Mirarab et al. 2014), which may give an overestimation of internode pairs in the anomaly zone. Second, we estimated branch lengths using BP&P v2.1b (Yang and Rannala 2010) with the original sequence data for 429 loci and the eMRC topology from MP-EST. BP&P uses a fixed species tree topology and the multispecies coalescent model of gene trees estimated using the Jukes–Cantor model to estimate branch lengths (t) and population sizes (θ) (Rannala and Yang 2003; Burgess and Yang 2008). A gamma prior on θ (a = 2.0, β = 200) with a mean of 0.01 was used for population size estimates on nodes. A gamma prior on t (a = 4.0, β = 10.0) was used for the root node height with other times generated from the Dirichlet distribution (Yang and Rannala 2010). Rate variation between loci was accommodated with the random-rates model (Burgess and Yang 2008), in which the average rate for all loci is fixed at 1 and the rates among loci are generated from a Dirichlet distribution. We used an α of 2.0 for moderate variation among loci. The Markov chain Monte Carlo was run for 100,000 samples, sampling every 10 generations for a total of 1,000,000 sampled states with a burnin of 150,000 states. Three independent analyses were conducted to verify convergence on a stable posterior. BP&P results were converted to coalescent units (λ = 2t/θ) consistent with those calculated by MP-EST.
Each pair of parent–child internodes were compared with the anomaly zone based on values of κ, calculated from BP&P and MP-EST. The value for the parent nodes (κ-nodes) were put into equation (1) for the limit of the anomaly zone in a four-taxon asymmetric tree to determine whether they are inside the zone and would therefore produce an AN. If the value of the child (κ-node) is less than κ(n), the pair of internodes are in the anomaly zone and AGTs are expected. This calculation was first conducted on the median values of branch lengths from the eMRC tree of the MP-EST species tree replicates and the median values of κ and θ from the eMRC tree of the BP&P posterior distribution. Additionally, for the MP-EST bootstrap replicates, anomaly zone calculations were done for each internode pair for each species tree bootstrap replicate, accounting for topological error in estimates of branch lengths. For BP&P, 1000 random draws of joint values of κ and θ for the internode pairs were made from the posterior distribution and compared with the anomaly zone. The BP&P analysis is only performed on the eMRC topology. Scripts (https://github.com/cwlinkem/anomaly_zone.git) to perform these functions relied on the Dendropy package (Sukumaran and Holder 2010). We report the proportion of bootstrap replicates that match the eMRC tree and are in the anomaly zone with MP-EST and the proportion of the 1000 draws from BP&P posterior distribution that are in the anomaly zone.

**RESULTS**

**Genomic Data and Assemblies**

All samples were successfully sequenced at sufficient levels to result in high coverage of target loci. Some samples represented a larger portion of the sequencing (Table 1) potentially due to unequal pooling prior to hybridization. Raw reads averaged 9.8 million (range of 5.4–21.6) reads per species with most reads being of high quality, resulting in a low rate of trimming and removal. Sequencing resulted in higher coverage than needed for the sequence-capture approach due to using 24 samples on the sequencing lane instead of the full potential of 96 samples. The added sequence coverage resulted in a higher proportion of off-target sequencing than would be expected, including complete mitochondrial genomes for many taxa. Results from other experiments have shown that off-target sequencing is reduced when more samples are included in the sequencing lane without a loss of target sequence coverage (unpublished data).

The 429 loci from the BP&P assembly totals 276,480 base pairs with 5.28% missing data and an average length of 644 base pairs (range of 338–1070). Individual loci vary in character variability with an average of 6% parsimony-informative sites (range of 0–15%) and 18% variable sites (range of 1–48%). Most loci

<table>
<thead>
<tr>
<th>Species</th>
<th>Voucher (million)</th>
<th>Total reads (million)</th>
<th>Clean reads (million)</th>
<th>Contigs</th>
<th>Loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branchiosaurus bontae</td>
<td>AJB075</td>
<td>8.71</td>
<td>7.71</td>
<td>29.942</td>
<td>550</td>
</tr>
<tr>
<td>Chalcides ocellatus</td>
<td>MVZ242790</td>
<td>6.59</td>
<td>5.93</td>
<td>26.745</td>
<td>569</td>
</tr>
<tr>
<td>Emia caeruleocauda</td>
<td>KU307155</td>
<td>6.08</td>
<td>5.64</td>
<td>21.414</td>
<td>563</td>
</tr>
<tr>
<td>Eurypleps taniotatus</td>
<td>MVZ246017</td>
<td>7.42</td>
<td>6.72</td>
<td>30.163</td>
<td>572</td>
</tr>
<tr>
<td>Labula elegans</td>
<td>BPRM16690</td>
<td>10.2</td>
<td>9.44</td>
<td>49.859</td>
<td>556</td>
</tr>
<tr>
<td>Lygosoma brevicauda</td>
<td>MVZ249721</td>
<td>8.40</td>
<td>7.70</td>
<td>47.774</td>
<td>553</td>
</tr>
<tr>
<td>Mahuya unimarginata</td>
<td>CWL613</td>
<td>19.5</td>
<td>17.9</td>
<td>134.666</td>
<td>533</td>
</tr>
<tr>
<td>Mesoscincus manguae</td>
<td>CWL614</td>
<td>11.4</td>
<td>10.5</td>
<td>38.828</td>
<td>550</td>
</tr>
<tr>
<td>Ophiomorus raithmai</td>
<td>MVZ248453</td>
<td>7.78</td>
<td>7.00</td>
<td>35.506</td>
<td>568</td>
</tr>
<tr>
<td>Plestodon fasciatus</td>
<td>KU389646</td>
<td>12.6</td>
<td>11.7</td>
<td>67.908</td>
<td>542</td>
</tr>
<tr>
<td>Scincus scincus</td>
<td>MVZ234538</td>
<td>6.54</td>
<td>5.94</td>
<td>20.231</td>
<td>566</td>
</tr>
<tr>
<td>Sphenomorphus tridigitus</td>
<td>FMNH258830</td>
<td>8.39</td>
<td>7.82</td>
<td>39.350</td>
<td>560</td>
</tr>
<tr>
<td>Sphenomorphus variatus</td>
<td>KU13087</td>
<td>8.99</td>
<td>8.31</td>
<td>43.927</td>
<td>550</td>
</tr>
<tr>
<td>Tylbosaurus sp.</td>
<td>MVZ248900</td>
<td>8.35</td>
<td>7.60</td>
<td>31.261</td>
<td>549</td>
</tr>
<tr>
<td>Tylbosaurus parvus</td>
<td>JAM62793</td>
<td>5.41</td>
<td>5.00</td>
<td>18.850</td>
<td>568</td>
</tr>
<tr>
<td>Xantusia vigilis</td>
<td>KU220092</td>
<td>21.3</td>
<td>19.9</td>
<td>155.978</td>
<td>550</td>
</tr>
</tbody>
</table>

* Iterative de Bruijn graph short-read assembler
* Museum of Vertebrate Zoology, UC Berkeley, CA.
* University of Kansas, Lawrence KS.
* Bernice Pauahi Bishop Museum, Honolulu, HI.
* No voucher specimen. Tissue deposited at KU.
* Field Museum of Natural History, Chicago IL.

**Loci Informativeness and Model Choice**

The 429 loci from the IDBA assembly totals 276,480 nucleotide positions with 5.28% missing data and an average length of 644 base pairs (range of 338–1070). Individual loci vary in character variability with an average of 6% parsimony-informative sites (range of 0–15%) and 18% variable sites (range of 1–48%). Most loci
have a best-fit model matching HKY or K80 (400 out of 429) suggesting a prevalence of transition/transversion bias in these genomic loci (Table 2). Model testing shows a preference for either a gamma or invariant-sites model (376 out of 429), but the combination is rarely preferred (38 out of 429). Only 18 loci have a preferred model that does not accommodate among-site rate heterogeneity (Table 2).

### Gene Trees

Maximum-likelihood (ML) searches for individual loci resulted in 429 unique topologies, one for each locus. These topologies also differ from the concatenation ML tree and the species tree. Bootstrap replicates average 925 (range of 163–1000) unique topologies out of the 1000 replicates indicating that gene tree resolution is low for most individual loci. Concatenated gene trees are largely congruent between the ML and Bayesian runs (Fig. 2). Acontinae is sister to all other taxa with strong support (100 bootstrap and 100 posterior probability). Lygosominae is monophyletic, and the Sphenomorphus group is sister to the clade of the Mabuya, Lygosoma, and Eugongylus groups. The Eugongylus group is sister to the Lygosoma group. All of these relationships have strong support in both analyses. There is also strong support for Brachymeles to be sister to Lygosominae, making Scincinae paraphyletic. The only topological difference between the two analyses is the placement of Ophiomorus. In the ML analysis, Ophiomorus is sister to all Scincinae and Lygosominae with a posterior probability of 71. In both cases, there is strong support for this placement. Ophiomorus is low despite the large amount of data used in these analyses.

### Species Trees

There were 110 unique species tree topologies found across the replicate species trees. The MP-EST eMRC species tree is 100% consistent for many relationships (Fig. 2c), but some key relationships vary among species tree replicates and differ from the concatenation tree. Most significantly, the placement of Brachymes in the species trees recovers a monophyletic Scincinae, but with moderate (77% of replicates) frequency. The alternative placement of Brachymes sister to Lygosominae (as seen in the concatenation analyses) has a frequency of 21%. The sister relationships of Scincus and Mesoscincus has the lowest frequency (48% of replicates). Ophiomorus is sister to Plestiodon with low frequency (53% of replicates). The Lygosominae portion of the species tree is identical to the concatenation topologies. The concatenation topology from RAxML, which had >99% support for all but one node, only matches 3% of the species tree replicates.

### Nodes in the Anomaly Zone

**MP-EST.**—Median internode lengths calculated with sumtrees (Sukumaran and Holder 2010) were used to calculate the anomaly zone for each pair of internodes in the eMRC (equation (1)). A large region of the phylogeny is in the anomaly zone based on these median values (Fig. 3a). The majority of relationships in Scincinae have internode lengths that are expected to produce AGTs. Examining the anomaly zone across the species tree replicates (Fig. 3b) shows that most pairs of parent–child internodes found to be in the anomaly zone with median branch lengths remain in the anomaly zone across replicates. One region of the phylogeny leading to Lygosominae (yellow branches) is only in the anomaly zone for 6% of the replicates that match the eMRC topology. The species tree replicates include relationships not found in the eMRC that represent the alternate resolutions of the poorly resolved nodes. Many of these alternative species tree relationships in the replicate set of trees are also in the anomaly zone (results not shown).

**BP&P.**—The posterior distribution from BP&P was summarized using sumtrees to obtain the median branch length $\tau$ and population size $\theta$ for internodes. BP&P calculates $\tau$ and $\theta$ individually, which were used to calculate $\lambda$ ($\lambda = 2\tau/\theta$) (Table 3). Internodes B, C, and N (Fig. 2c) are particularly short, and internode G is the longest. Calculations of $a(x)$ (equation (1)) show that six internodes (B, C, J, K, L, N) have $a(x)$ values above zero, indicating that they may produce AGTs depending on the length of the descendant internode. Of these, B, C, J, and L have descendant internodes. The internode pairs C/J, B/L, and K/N are in the anomaly zone because the lengths of internodes C and I are shorter than $a(x)$ for B and internode N is shorter than $a(x)$ for L (Fig. 3c). To account for the range in branch length estimates across
FIGURE 2. Majority-rule consensus trees from the concatenated loci run in RAxML (a), ExaBayes, (b) and the species tree from MP-EST (c). Open circles indicate 100% bootstrap support or a Bayesian posterior probability of 1.0. Nodes with lower support are labeled. Both concatenation analyses have similar topologies with differences in the placement of Ophiomorus, and the RAxML tree is slightly longer. The species tree (c) is shown as a cladogram. Relationships within Lygosominae are the same across analyses. The relationships within Scincinae differ both between the concatenation analyses and in comparison with the species tree. Letters on species tree nodes are used for Table 3 and discussion of internode pairs.
FIGURE 3. Majority-rule consensus topology shown as a cladogram. Pairs of internodes that are under the anomaly zone curve are highlighted in bold on each tree. Tree (a) is based on median values of branch lengths calculated from MP-EST. Tree (b) shows the frequency of the internodes in the anomaly zone across species tree replicates that match the EMRC topology. Tree (c) shows the occurrence based on the median values of branch lengths from the posterior distribution of BP&P. Tree (d) shows the frequency of the internode in the anomaly zone based on 1000 draws of joint internode values from the posterior distribution.

TABLE 3. Median branch lengths (τ) and population sizes (θ) used to calculate the branch length in coalescent units (λ) for the internodes labeled on the MP-EST species tree (Fig. 2c).

<table>
<thead>
<tr>
<th>Internode</th>
<th>τ</th>
<th>θ</th>
<th>λ</th>
<th>a(τ)</th>
<th>Descendant internode(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.006391</td>
<td>0.016252</td>
<td>0.781678</td>
<td>&lt; 0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0.000306</td>
<td>0.012839</td>
<td>0.003778</td>
<td>1.002</td>
<td>C, I</td>
</tr>
<tr>
<td>C</td>
<td>0.00117</td>
<td>0.008111</td>
<td>0.293856</td>
<td>0.443</td>
<td>D, E</td>
</tr>
<tr>
<td>D</td>
<td>0.01574</td>
<td>0.06686</td>
<td>0.45899</td>
<td>&lt; 0</td>
<td>G</td>
</tr>
<tr>
<td>E</td>
<td>0.002352</td>
<td>0.007271</td>
<td>0.701866</td>
<td>&lt; 0</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>0.015353</td>
<td>0.001035</td>
<td>0.309054</td>
<td>&lt; 0</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>0.007712</td>
<td>0.007601</td>
<td>2.02657</td>
<td>&lt; 0</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>0.002186</td>
<td>0.00518</td>
<td>0.49652</td>
<td>&lt; 0</td>
<td>I</td>
</tr>
<tr>
<td>I</td>
<td>0.005576</td>
<td>0.014087</td>
<td>0.79165</td>
<td>&lt; 0</td>
<td>J</td>
</tr>
<tr>
<td>J</td>
<td>0.00754</td>
<td>0.024723</td>
<td>0.34189</td>
<td>0.0032</td>
<td>L, K</td>
</tr>
<tr>
<td>K</td>
<td>0.000902</td>
<td>0.019889</td>
<td>0.128958</td>
<td>0.0982</td>
<td>M, N</td>
</tr>
<tr>
<td>L</td>
<td>0.000588</td>
<td>0.010911</td>
<td>0.10778</td>
<td>0.1284</td>
<td>M, N</td>
</tr>
<tr>
<td>M</td>
<td>0.002272</td>
<td>0.00599</td>
<td>0.26527</td>
<td>&lt; 0</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>0.000354</td>
<td>0.012777</td>
<td>0.024105</td>
<td>0.4994</td>
<td>—</td>
</tr>
</tbody>
</table>

a(τ) indicates the limit of the anomaly zone for that internode length. If the length of the descendant internode is smaller than a(τ), then the internode pair is in the anomaly zone. If a(τ) = < 0 the internode pair will not produce anomalous trees. Descendant internodes in bold indicate anomaly zone pair.

The Anomaly Zone in Empirical Phylogenies

Species trees estimated using coalescent methods often differ from concatenated gene trees, and the anomaly zone may be contributing to these disagreements. The anomaly zone could also explain why some phylogenomic studies find low support for relationships despite the inclusion of hundreds to thousands of loci. In the empirical example of skinks presented here, we find strong conflict between species trees and concatenated gene trees, as well as conflict between individual gene trees. This type of conflict is typical for phylogenomic studies of rapid diversification events (Zou et al. 2008; McCormack et al. 2012; Jarvis et al. 2014). Our examination of the anomaly zone in skinks shows that the parts of the tree in conflict correspond with areas of the tree that are also estimated to be in the anomaly zone. The anomaly zone is a potential explanation for why conflicting relationships persist in some phylogenomic studies.
interested in whether or not we can estimate individual gene trees. The species tree is the target of analysis, not the individual gene trees (Edwards 2009). While it is true that estimation of any particular gene tree is hindered by low genetic variation, this does not change whether or not speciation events occurred quickly enough to place them into the anomaly zone. Phylogenetic analysis of concatenated loci, as a proxy for the species tree, can result in strong support as more loci are combined, even if there is low genetic variation in individual loci. If the speciation history is in the anomaly zone, then the resulting phylogeny will be erroneous (Kubatko and Degnan 2007). Our study shows that there is low support for relationships in most individual gene trees, but after concatenating the 276,480 characters they provide strong support for most relationships (Fig. 2a–b). Huang and Knowles (2009) showed that when the species history is in the anomaly zone, gene tree estimation error will be high and that the lack of variation will make estimation of individual gene trees difficult. The observation of significant gene tree discordance across hundreds of loci may be a good sign that the species history is in the anomaly zone, and the framework that we provide here offers one way to test this hypothesis.

Summary statistic methods for species tree inference have become a necessity, since the more statistically rigorous full Bayesian approaches (*BEAST and BEST) cannot handle hundreds of loci (Bayzid and Warnow 2013). Summary methods use gene trees to estimate species trees, and as previously discussed, gene tree estimation error is typically high in cases where an anomaly zone is suspected. Gene tree estimation error reduces the accuracy of species tree estimation (Bayzid and Warnow 2013; Mirarab et al. 2014), which makes species tree inference in the anomaly zone more difficult (Liu and Edwards 2009). We accounted for gene tree error by repeating the species tree estimation procedure using the gene trees constructed from the bootstrap replicates, which can provide a measure of accuracy that is not available when only using the ML gene trees (Mirarab et al. 2014). Despite this, it is still possible that gene tree estimation error is biasing the species tree. Ideally, these inaccuracies should be reflected by the low frequency of occurrence for inaccurate nodes in the species tree. By estimating the anomaly zone across all topologies in the replicate species trees from MP–EST, we find that the anomaly zone can be inferred even when species trees contain low supported nodes. Species trees in the anomaly zone are likely to have nodes with low frequency due to the high frequency of AGTs and the occurrence of gene tree estimation error. In these situations, including more loci may not increase node frequency due to the addition of more gene tree estimation error. Increasing the accuracy of gene trees through sampling longer loci (McCormack et al. 2009), or by combining loci with a shared history into larger gene fragments (Bayzid and Warnow 2013; Betancur-R et al. 2013), may improve inference of species trees in the anomaly zone. Inferring relationships in recently diverged species may also benefit from increased sampling of individuals within each species (Rokas and Carroll 2005; Maddison and Knowles 2006), though this benefit would be lost in deep divergences (Kubatko and Degnan 2007).

Frequency of AGTs

While we cannot be certain that the skink species tree that we estimated is correct, it is clear that the set of species tree topologies estimated from these phylogenomic data show signs of the presence of the anomaly zone over multiple pairs of internodes (Fig. 3). The extent and frequency of the anomaly zone in this empirical example indicates that many AGTs may exist across the genomes of these taxa. It is important to keep in mind that each inference of the anomaly zone is limited to pairs of internal branches, without consideration of neighboring relationships. Only considering neighboring internal branches is a simplification; accounting for other branches would increase the size of the anomaly zone (Rosenberg and Tao 2008). Focusing on pairs of internal branches makes our investigation of the anomaly zone a conservative approximation of the extent of AGTs. When considered together, the extent of the anomaly zone in our empirical example has the potential to produce many AGTs.

Rosenberg and Tao (2008) showed that the number of AGTs increases rapidly as the number of short internodes increases and that an anomaly zone that includes three internodes (five-taxon trees or larger) can produce as many as 45 AGTs. Estimates for trees larger than five taxa or an anomaly zone spanning more than three nodes have not been estimated, but are expected to increase exponentially. In the example of skinks, the 429 loci sampled are all estimated to have unique ML topologies. Our sample of loci does not show a trend for a particular topology to be more frequent, but this is not unexpected given the large number of possible topologies (4.22 × 1013) for a 16-taxon rooted tree. Across the 429 alternate topologies sampled, an eMRC of these topologies shows that most splits occur in low frequency. Splits associated with internode pairs estimated to be in the anomaly zone occur at lower frequency among the gene trees than splits not estimated in the anomaly zone. It is unclear how species tree methods perform when the anomaly zone is larger than a four-taxon case as we see in skinks.

Simulation studies testing the ability of different species tree methods to overcome the four-taxon anomaly zone (Kubatko and Degnan 2007; Huang and Knowles 2009; Liu and Edwards 2009) have shown that species tree methods can accurately estimate the true species tree when provided with enough data. For concatenated gene trees, we know that combining loci with conflicting topologies can result in strong support for a single topology (Rokas et al. 2003). In some cases, relationships within the concatenated gene tree may not occur within the individual gene trees, a result referred to as hidden likelihood support (Gatesy and Baker 2005).
This hidden support has been argued to be an advantage of concatenation methods over species tree methods (Springer and Gatesy 2014; Gatesy and Springer 2014). In cases where the speciation history fits the demographic scenario of the anomaly zone, the hidden likelihood support in the concatenated gene tree will be for an incorrect topology, resulting in strong inference of an incorrect species history (Kubatko and Degnan 2007). The method for inferring the anomaly zone presented here allows researchers to identify potential conflicts and decide on appropriate analytical solutions.

Edwards (2009) predicted that phylogenomic studies would find lower support for relationships using species tree methods than would be obtained from concatenation, especially in older clades. This prediction is based on the idea that missing data would have a larger effect on species trees and that species trees use more complex models compared with concatenation. We find these predictions to be accurate, though their cause may be different than originally proposed. We propose that low support in species tree analyses is due to a combination of gene tree estimation error and the presence of AGTs that bias phylogenetic signal. Simulation studies have shown that many species tree methods can overcome the anomaly zone in simple four-taxon scenarios (Liu and Edwards 2009), but no study has looked at the effect of larger anomaly zone problems on trees with more taxa. We predict that when the anomaly zone occurs across more than two internodes, the greater number of AGTs will provide support for multiple species trees, reducing the support for some species tree nodes. This will likely occur even when gene trees are estimated with certainty (i.e., using simulated data), a luxury not available with empirical studies. The increased number of AGTs may also result in low support in concatenation analyses even when analyzing hundreds of genes if the AGTs are in direct topological conflict. In this study, we find two nodes with *RAxML* and a different node in *EXaBayes* that have low support despite having over 276,000 (50,666 variable; 15,524 parsimony informative) characters in the analyses. With multiple AGTs there may be alternative topologies with high probability in the set of candidate genes. These alternative topologies should lower branch support in concatenation analyses. This low support for nodes near short branches may be an indication of an anomaly zone problem in phylogenomic data sets.

*Higher-Level Skink Relationships*

Resolving the relationships within Scincidae is an ongoing challenge (Brandley et al. 2012; Wiens et al. 2012; Pyron et al. 2013; Lambert et al. 2015) which our current study addresses with a slightly different approach compared with previous studies. Pyron et al. (2013) used hundreds of taxa in a sparse matrix, while Wiens et al. (2012) and Lambert et al. (2015) used many loci with few taxa. Our taxon sampling is most similar to Brandley et al. (2012) but utilizes 429 loci and species tree analyses to estimate relationships. Our preferred estimate of species relationships (Fig. 2c) shows subfamily relationships concordant with results in Pyron et al. (2013) and Lambert et al. (2015). Acontinae is sister to all other skinks, and the subfamilies Scincinae and Lygosominae are monophyletic. The monophyly of Scincinae conflicts with most studies and our own results from analyses of concatenated data (Whiting et al. 2003; Brandley et al. 2005, 2012; Wiens et al. 2012) but is likely the more accurate relationship based on our inference of the anomaly zone in relation to the nodes preceding and in this subfamily. Similar to Lambert et al. (2015), we find *Brachymeles* to be related to other Scincinae as opposed to sister to all Lygosominae (Brandley et al. 2012). The relationships among Scincinae genera differs from those presented in Brandley et al. (2012) though we have far fewer genera sampled and low support so many comparisons should be made with caution. We consider the alternative placement of *Brachymeles* sister to Lygosominae, seen in concatenated gene trees, to be an artifact of rapid diversification resulting in a series of AGTs. A detailed examination of the relationships among the genera in Scincinae with broad taxon sampling is clearly warranted. Our inference of the anomaly zone among some Scincinae genera suggests that hundreds of loci and species tree analyses will be necessary to accurately estimate the phylogenetic relationships within this group.

Relationships within Lygosominae are largely concordant with previous studies in finding the *Sphenomorphus* group sister to all other groups (Honda et al. 2003; Reeder 2003; Skinner et al. 2011). We find the *Lygosoma* group to be sister to the *Eugongylus* group and that this pair is sister to the *Mabuya* group consistent with the results of Reeder (2003) and Skinner et al. (2011). Relationships among the sampled genera in the *Sphenomorphus* group are consistent with previous results (Linkem et al. 2011). Over three-quarters of all skink species are in Lygosominae, and it appears so based on our limited sampling that the broad groupings of genera can be consistently resolved and that the anomaly zone is not an issue at this level. Within the *Sphenomorphus* group, previous studies have reported short branches separating major groups (Linkem et al. 2011; Skinner et al. 2011). We suspect that there will be anomaly zone issues within the *Sphenomorphus* radiation, especially in Australia. The *Eugongylus* group will likely also present anomaly zone issues given the large number of species in the group and relatively recent origin. Studies at a more fine scale will likely find similar rapid diversification patterns that are consistent with the anomaly zone. Further work on these diverse groups is needed to better understand their systematic relationships.

A monophyletic Scincinae necessitates an investigation into morphological features to define this group. Greer (1970) considered Scincinae to be the primitive morphological form from which all other skinks were derived, but molecular studies find Scincinae taxa to be nested and sister to Lygosominae,
conflicting with the prevailing morphological interpretation. A detailed and rigorous analysis of the morphological evolution and biogeography of skinks will require much more rigorous taxon sampling.

Overcoming the Anomaly Zone

Sequencing hundreds of unlinked single copy loci provides an opportunity to explore the conflicts between loci and analytical approaches, as well as address what may be the source of conflict. Our work shows that researchers conducting empirical studies should closely consider the potential impact the anomaly zone has on their phylogenetic analyses. A common trend with phylogenomic studies is to analyze the data with concatenation, an approach that has the advantage of faster computation times and simplicity, but that provides overwhelmingly and likely erroneous strong support across most of the tree. Species tree analyses often result in lower support for difficult parts of the tree than concatenation. Instead of marginalizing the species tree results, we should acknowledge that they are a likely consequence of the speciation history for the group. The lower support provided by coalescent-based species tree inference is potentially a more accurate reflection of the support for the tree given the data.

As we show here, the anomaly zone is likely more pervasive than previously suggested and should be accounted for when studying taxa that may have diverged rapidly, even if that rapid event was in the distant past. Combining hundreds to thousands of independent loci together with coalescent-based species tree inference is the most effective way of getting an accurate result. Targeting longer loci will help reduce the number of faster computation times and simplicity, but that provides overwhelmingly and likely erroneous strong support across most of the tree. Species tree analyses often result in lower support for difficult parts of the tree than concatenation. Instead of marginalizing the species tree results, we should acknowledge that they are a likely consequence of the speciation history for the group. The lower support provided by coalescent-based species tree inference is potentially a more accurate reflection of the support for the tree given the data.

As we show here, the anomaly zone is likely more pervasive than previously suggested and should be accounted for when studying taxa that may have diverged rapidly, even if that rapid event was in the distant past. Combining hundreds to thousands of independent loci together with coalescent-based species tree inference is the most effective way of getting an accurate result. Targeting longer loci will help reduce the number of faster computation times and simplicity, but that provides overwhelmingly and likely erroneous strong support across most of the tree. Species tree analyses often result in lower support for difficult parts of the tree than concatenation. Instead of marginalizing the species tree results, we should acknowledge that they are a likely consequence of the speciation history for the group. The lower support provided by coalescent-based species tree inference is potentially a more accurate reflection of the support for the tree given the data.

As we show here, the anomaly zone is likely more pervasive than previously suggested and should be accounted for when studying taxa that may have diverged rapidly, even if that rapid event was in the distant past. Combining hundreds to thousands of independent loci together with coalescent-based species tree inference is the most effective way of getting an accurate result. Targeting longer loci will help reduce the number of faster computation times and simplicity, but that provides overwhelmingly and likely erroneous strong support across most of the tree. Species tree analyses often result in lower support for difficult parts of the tree than concatenation. Instead of marginalizing the species tree results, we should acknowledge that they are a likely consequence of the speciation history for the group. The lower support provided by coalescent-based species tree inference is potentially a more accurate reflection of the support for the tree given the data.

As we show here, the anomaly zone is likely more pervasive than previously suggested and should be accounted for when studying taxa that may have diverged rapidly, even if that rapid event was in the distant past. Combining hundreds to thousands of independent loci together with coalescent-based species tree inference is the most effective way of getting an accurate result. Targeting longer loci will help reduce the number of faster computation times and simplicity, but that provides overwhelmingly and likely erroneous strong support across most of the tree. Species tree analyses often result in lower support for difficult parts of the tree than concatenation. Instead of marginalizing the species tree results, we should acknowledge that they are a likely consequence of the speciation history for the group. The lower support provided by coalescent-based species tree inference is potentially a more accurate reflection of the support for the tree given the data.


