UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Assessing and comparing applications of molecular clocks to phylogenetic datasets

Leandro Rodolfo dos Santos Ribeiro

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Professor Octávio S. Paulo

2020

Resumo

A implementação de relógios moleculares na inferência filogenética foi uma dos maiores descobertas no estudo da área, permitindo à comunidade científica obter resultados mais robustos e com uma menor margem de erro significativa, suportados pelo uso de registos fósseis e geológicos.
Nesta tese foram comparados os métodos de inferência filogenética molecular implementados no programa *BEAST2* e *StarBEAST2,* para que se tenha uma ideia de em que cenário um é mais vantajoso aplicar um ou o outro assim como a comparação entre as suas versões mais recentes e antigas.
Assim sendo, foram realizadas inferências moleculares independentes a espécies de escaravelhos tigre (Coleoptera: Cicindelini) e a espécies de escaravelhos subterrâneos (*Trechus fulvus*). As análises tiveram por base a aplicação de um *workflow* que permite a realização de inferências filogenéticas moleculares inter e intraespecíficas, com base nos recursos disponibilizados pelo pacote do programa *BEAST2.*
Os resultados obtidos com o template standard de BEAST2 demonstram que é fundamental o uso de um *outgroup* adequado na calibração da árvore filogenética, que garanta suporte de confiança à inferência realizada. Relógios inapropriados ou calibrações incorretas (resultando em valores mínimos de ESS) conduzem a resultados ilusórios de escalas de tempo.
Contudo com o programa *StarBEAST2* não é necessário o uso de um *outgroup, assim como* demonstra obter resultados com melhor resolução, intervalos de divergência menores e um tempo de processamento significamente mais rápido comparativamente ao programa *BEAST2* (dependendo dos specs do PC usado) ao mesmo tempo usando um método de coalescência multi-espécies o qual habilita ao programa de modelar as árvores de genes dentro da árvore de espécies resultando em geral em resultados de tempos de divergência mais precisos. Porém deve-se ter sempre em atenção o contexto biológico ao averiguar os resultados obtidos na inferência, pois mesmo que sejam bem suportados pela análise dos valores de confiança, nunca deverão ser interpretados isoladamente.

Palavras Chave

Inferência filogenética; Relógios Moleculares, *BEAST2*; *StarBEAST2;* MultiSpecies Coalescent Process

Summary

The implementation of molecular clocks in phylogenetic inference was one of the largest breakthroughs in the field of genetics and phylogeny, allowing to the scientific community a way to obtain more robust results and with a significantly smaller margin of error, supported by the use of fossil and geological records.

In this thesis were compared the methods of molecular phylogenetic inference in the programa BEAST2 and *BEAST2 (or starBEAST2), so as to identify in which scenario is better to use one or the other as well as comparing their most recent versions to their previous iterations so as to ascertain how much better they've gotten if at all.

And so, were realized phylogenetic inferences on species of tiger beetles (Coleoptera: Cicindelini) and on species of ground beetles (*Trechus fulvus*). The subsequent analyses had as their core the usage of a workflow designed to standardize the whole procedure permitting the realization of interspecific and intraspecific molecular phylogenetic inferences based on the available resources in BEAST2's packages.

The results obtained through the standard template of BEAST2 show that the usage of an adequate outgroup tailored to the calibration of the phylogenetic tree is fundamental to the credibility of the inference in question. Inaccurate clocks or incorrect calibrations (resulting in low ESS values) inevitably lead to illusory results.

However, with the usage of StarBEAST2 an outgroup isn't required to the phylogenetic inference, as well as it showing (normally) results with better resolution, smaller divergence times and a much faster processing time when compared to BEAST2 (depending on the PC specs) all the while using a multi species Coalescent methods that allows the program to model the analyzed gene trees within the species tree thereby improving the overall accuracy and precision of the divergence times. Although one should always take the biological context into account when analyzing the results as they never should be treated immediately as fact even if well supported by its confidence values.

Keywords

Phylogenetic Inference; Molecular Clocks, *BEAST2*; *StarBEAST2;* MultiSpecies Coalescent Process

Índice

Índice de Figuras

Índice de Tabelas

Lista de Siglas e Abreviaturas

API - *Application programming interface*

BEAST - *Bayesian Evolutionary Analysis Sampling Trees*

COI - *Cytochrome c Oxidase Subunidade I*

Entrez - *Global Query Cross-Database Search System*

ESS - *Effective Sample Size*

LS - *Least Square*

Ma - Milhão de anos

MAS - *Multiple sequence alignment*

MCMC - *Markov chain – Monte Carlo*

ME - *Minimum Evolution*

ML - *Maximum Likelihood*

MP- *Maximum Parsimony*

MRCA - *Most Recent Common Ancestral*

NCBI - *National Center for Biotechnology Information*

NJ - *Neighbour-Joining*

OTU - *Operational Taxonomic Unit*

PP - probabilidade posterior

UPGMA - *Unweighted Pair Group Method with Arithmetic Mean*

1 Introdução

1.1 Objetivos

BEAST2 e *BEAST2 são das aplicações mais conhecidas quando se trata do uso de relógios moleculares para a construção de árvores filogenéticas e por conseguinte a sua inferência. Este Projecto visa comparar diferentes publicações com o uso destes programas com o intuito de averiguar qual o mais indicado para cada tipo de situação assim como as suas vantagens e desvantagens.

Adicionalmente também se irá reanalisar os artigos escolhidos com as versões mais recentes e antigas de ambos os métodos de modo a verificar qual ou quais são os principais melhoramentos com cada iteração e se irá influenciar a escolha do método a usar.

1.2 Tarefas

Tarefa A – Aprendizagem dos programa *BEAST2* e **BEAST2*: leitura e escolha das publicações adequadas ao projeto, leitura da documentação do programa; reprodução de inferências a partir de exemplos;

Tarefa B – Inferência filogenética molecular em espécies de escaravelhos tigre (Coleoptera: Cicindelini), com os métodos *BEAST2* e **BEAST2* baseada no gene *COI;*

Tarefa C – Inferência filogenética molecular em escaravelhos do grupo *Trechus fulvus*, com os métodos *BEAST2* e **BEAST2* a partir do gene *COI*;

Tarefa D – Inferência filogenética molecular em escaravelhos tigre (Coleoptera: Cicindelini), com o programa *BEAST2* e **BEAST2* a partir dos genes *16S e COI* (*dataset* concatenado)*;*

Tarefa E – Inferência filogenética molecular em escaravelhos do grupo *Trechus fulvus*, com o programa *BEAST2* e **BEAST2* a partir dos genes *COI, LSU, SSU e rrnL+trnL+nad1* (*dataset* concatenado)*;*

Tarefa F – Comparação de resultados entre Inferências realizadas com versões anteriores do programa *BEAST2 (v.2.4.0);*

2 Fundamentos teóricos

Vários biólogos na comunidade científica concordam que o estudo de árvores filogenéticas deve ser um dos alicerces em investigações em várias áreas da Biologia visto o estudo destas providenciar novos e detalhados conhecimentos na área, esta realização é agora aparente a investigadores de diversas campos incluindo a ecologia, biologia molecular e fisiologia. Um exemplo do valor destes estudos é o melhor entendimento do funcionamento do processo evolutivo proporcionando a investigadores e indústrias a informação relevante que por conseguinte poderá ser usada para o estudo genómico e melhoramento da qualidade de produtos[1, 2].

Para o propósito deste projecto vamos focar-nos nos métodos desenvolvidos que visam averiguar a recriação da história evolutiva das espécies usando sequências de, por exemplo, mtDNA[1, 2].

2.1 Filogenia

A reconstrução da história evolucionária de genes e espécies, a partir de métodos estimativos, é actualmente um dos temas mais importantes em Evolução molecular, A partir da relação entre *OTUs* (*Operacional Taxonomic Units*) (P. Ajawatanawong, 2017) a relação filogenética de organismos é apresentada em diagramas em forma de árvore denominadas árvores filogenéticas.

Através da inferência filogenética para além de ficarmos a saber como as sequências estudadas chegaram ao estado em que encontram hoje também nos proporciona um meio de como podem vir a alterar no futuro tornando a inferência filogenética fundamental em várias áreas de investigação, tais como o estudo da sistemática biológica e da biodiversidade (C. Senés-Guerrero et al., 2014), epidemiologia molecular (E. Kenah et al., 2016), identificação de funções genéticas (A. B. Chang et al., 2004), estudos do microbioma (J. B. H. Martiny et al., 2015), análises forenses (M. Siljic et al., 2017), descoberta de novos fármacos (K. A. Jacobson et al., 2014), e até mesmo no desenvolvimento de vacinas (S. Ojosnegros e N. Beerenwinkel, 2010). (P. Ajawatanawong, 2017)

2.1.1 Análise filogenética com base em dados de sequências

Após a aquisição de sequências de nucleótidos, em formato FASTA (embora haja outros formatos o formato FASTA é o mais comum nestes ensaios) é necessário como primeiro passo o uso de métodos de modo a formatar as sequências a usar a que damos o nome de alinhamento. Em seguida é criado um ficheiro FASTA onde possuímos os alinhamentos de todos os organismos permite dar resposta aos programas de filogenética, a partir do qual é possível obter um *input* matricial que pode ser modelado noutros formatos como o *Nexus*. (P. Ajawatanawong, 2017)

2.1.2 Árvores Filogenéticas

Com os resultados de inferência filogenética podemos representar estes mesmos com gráficos em esquema de árvore. A interpretação destes gráficos são então úteis na inferência das relações evolutivas num grupo de *OTUs*. Na maioria dos casos, o *OTU* representa uma espécie, mas também pode representar individualmente organismos de uma população, sequências de genes e proteínas ou um táxon, independentemente do *rank* taxonómico (família, ordem, classe, filo) (P. Ajawatanawong, 2017) no caso deste projecto, sequências de genes mitocondriais.

Por norma usa-se na construção das árvores uma abordagem de topo para a base, aos nódulos situados no extremo da árvore dá se o nome de nódulos externos, que representam cada uma das *OTUs* envolvidas na análise enquanto o ancestral hipotético mais recente entre duas *OTUs* é designado de nódulo interno. Assim, a ligação que se dá entre estes dois tipos de nódulos é denominada por ramo, sobre a qual se demonstra a relação evolutiva entre os taxa usados. O ramo que liga dois nódulos internos é classificado como ramo interno, sendo caracterizado como uma relação antiga. Da mesma forma, o ramo que liga um nódulo interno a um nódulo externo é denominado por ramo externo. A ou raiz representa a primeira divergência da árvore, o mais recente ancestral em comum (MRCA) de todos os taxa / *OTU's* ou nódulos externos*;* (P. Ajawatanawong, 2017). A *Figura 1* exemplifica a composição de uma árvore filogenética, bem como os principais termos acima referidos.

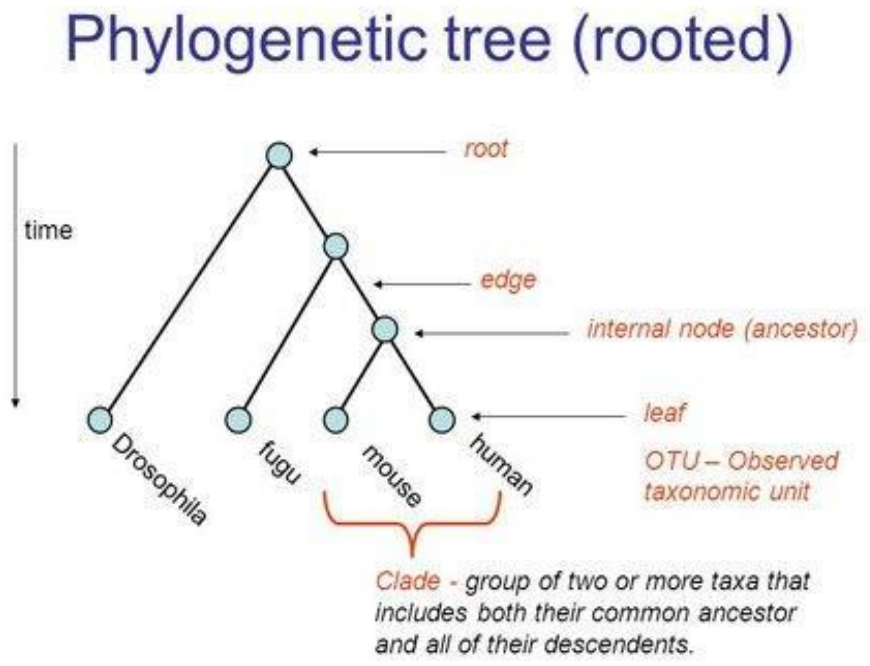


Figura 1 - árvore filogenética; Fonte (P. Ajawatanawong, 2017).

Por norma o programa BEAST para a construção de uma árvore filogenética apenas consegue reconstruir uma árvore sem raiz. De maneira a conferir um maior valor ao diagrama num contexto evolutivo, é preferível inferir uma árvore enraizada a partir da identificação da origem de todos os taxa. Para isto adiciona-se um *Outgroup ao dataset* (adicionando-se um prior monofilético contendo o grupo de organismos mais distante na partição ao ficheiro xml) (P. Ajawatanawong, 2017). O *outgroup* é geralmente um *OTU, taxon* ou grupo de organismos mais distante das espécies a serem estudadas servindo como ponto de referência na aquisição do ponto de origem das relações evolutivas do *ingroup* (que é composto pelos taxa do dataset a ser estudado) de modo a calibrar as relações da árvore. Tecnicamente com este método a *root* da árvore está assim localizada entre o *outgroup* e os restantes taxa. Com esta informação em mente, um bom *outgroup* será um *OTU* que tenha divergido a uma altura relativamente recente aos restantes taxa, mas suficientemente distinto dos mesmos a ponto de não inferir com o *ingroup.* (P. Ajawatanawong, 2017)

Todos os taxa que descendem do mesmo ancestral em comum são definidos como grupo monofilético ou *clade*. Já a um grupo de *OTU* que partilham o mesmo ancestral, mas que não possuem todos os membros descendentes, dá se o nome de grupo parafilético ou *glade* (P. Ajawatanawong, 2017).

2.1.3 Métodos de Inferência Filogenética

Dentro do métodos de inferência filogenética temos como principais blocos ou partes os seguintes:

- Critérios de optimização, que como o nome indica, visam maximizar e/ou minimizar os valores dos parâmetros introduzidos, proporcionando uma base de comparação entre árvores (estes serão abordados na próxima secção).
- Avaliação das árvores, usando os critérios referidos acima é seleccionada a melhor árvore (P. Ajawatanawong, 2017).
- Modelos evolutivos, tal como os critérios de optimização os modelos evolutivos visam exactamente o que se espera, usar testes de hipóteses de modo a modelar a evolução na Inferência.

2.1.3.1 Métodos de Distâncias

Uma das vertentes da secção anterior, estes métodos têm como por base do seu procedimento duas fases, a primeira destas implica o cálculo da distância evolucionária (medida da diferença de material genético entre diferentes espécies/individuos) para cada par de sequências a usar. Esta informação é de seguida guardada e formatada numa matriz de distâncias (P. Ajawatanawong, 2017) a partir de uma matriz de alinhamentos convertendo o emparelhamento das sequências em valores de distância (dissimilaridades).

Assim que esta matriz é originada, a matriz de alinhamentos usada para a criação desta é descartada pois para a finalidade deste processo a matriz de alinhamentos não será mais útil no realização da inferência filogenética. Existem vários métodos que possibilitam inferir qual a melhor árvore como

a UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*) (H. A. Khan et al., 2008), *Neighbour-Joining* (NJ, consultar *Tabela 1*) (N. Saitou e M. Nei, 1987), *Least Square* (LS) (J. Wen et al., 2018) ou *Minimum Evolution* (ME) (S. Bastkowski et al., 2016); (P. Ajawatanawong, 2017).

A característica que coloca os métodos de distâncias acima em relação a outros métodos é a superior velocidade de processamento, esta elevada velocidade é devido ao uso da matriz de distâncias o que implica a redução de informação no processo de conversão a matriz, embora esta seja uma vantagem em termos de velocidade a redução de informação implica uma perda de informação genética o que pode tornar esta opção menos desejável em casos de homosplasia (D. Ortega-Del Vecchyo et al., 2017) isto é, o surgimento de características semelhantes entre duas ou mais espécies, que não possuam proximidade genética entre si, este problema pode ser mitigado se a homosplasia for rara. Porém para os casos em que temos um número elevado de sequências semelhantes um dos métodos desta categoria é especificamente usado para tais casos, sendo este o NJ (Neighbour joining) .(P. Ajawatanawong, 2017)

2.1.3.2 Métodos de Caracteres

Passando agora dos métodos de distâncias para os métodos de Caracteres, estes são mais indicados para o uso com caracteres, como por exemplo, sequências de DNA assim como proteínas. Como estes métodos não convertem a informação dos caracteres em matrizes ou qualquer tipo de alteração isto torna estes métodos intrisicamente mais lentos que os métodos de distâncias contudo como não há perda de informação estes métodos acabam por ser significativamente mais precisos. Exemplos deste métodos contém a *Maximum Parsimony* (MP) (C. B. Stewart, 1993), a *Maximum Likelihood* (ML) assim como métodos de inferência Bayesiana. Em termos simples, o algoritmo destes métodos começa por fazer o *score* a todas as possibilidades filogenéticas a partir de *n* taxa. Quanto maior o score melhor será a árvore resultante, considerando os dados das sequências. (P. Ajawatanawong, 2017)

Método	Descrição	Vantagens	Desvantagens
Métodos de construção de árvores filogenéticas			
"Stepwise addiction"	Contrói uma árvore completa, começando com três sequências e adicionando incrementlmente novas sequências, uma de cada vez, à branch que possui a árvore mais adquada.	Método de rápida implementação; Etapas posteriores podem reverter processos anteriores.	Processa uma árvore, que muitas vezes não é a mais indicada; sequências alternativas adicionais podem contruir árvores com uma topologia diferente; não é tão rápido como o "neighbour-joining"
"Star decomposition"	Constrói uma "resolved tree", começando por todas as sequências estarem conectadas a um único 'hub node'. A cada passo, duas linhagens são adicionadas ao 'hub node', tornando-se "neighbours", escolhidos de forma a que a árvore seja mais adequada à inferência.	Método de rápida implementação; sequencias adicionais são irrelevantes	Processa uma árvore, que muitas vezes não é a mais indicada; os "neighbours" não podem ser separados em etapas posteriores; não é adquado a alguns métodos.
"Neigbour joining"	Um método de "Star decomposition" que usa a aproximação a um minimo evolutivo (critério de optimização)	Um dos métodos mais rápidos na construção de árvores evolutivas	Processa uma árvore, que muitas vezes não é a mais indicada; os "neighbours" não podem ser separados em etapas posteriores; não é adquado a alguns métodos.
Métodos de "search" a árvores filogenéticas			
"Heuristic search"	Fornece uma árvore i exaustiva :ontem todas as sequências, executa o processo de "branch swapping" para produzir árvores alternativas, de modo a encontrar uma árvore mais optimizada	Máis rápido que as "Exact searches"	Pode não encontrar a árvore mais adquada
"Exact search"	Processa uma search "exautiva" de modo a examinar todas as árvores possiveis, garantindo a melho como output. As técnicas de "branch-and-bound" podem eliminar as piores árvores continuando a garantir o melhor output.	O único método que garante inferir a melhor árvore	Tempo de processamento: apenas praticável para processar algumas sequencias (<20)

Tabela 1 - Métodos de construção e de search; Fonte (M. Holder e P. O. Lewis, 2003).

Os algoritmos de pesquisa (algoritmos que tomam um problema como entrada e retornam uma solução ao problema), dentro do campo da filogenia estes algoritmos são o branch-and-bound (E. L. Lawler e D. E. Wood, 1966)(que visa retornar uma soluções para problemas de optimização) e o método heurístico (A. Goëffon et al., 2010), já referidos acima (Tabela 1). O Algoritmo branch-and-bound tem como início usar três taxa do dataset escolhido na criação de uma árvore, continuando a sua construção aleatóriamente até o algoritmo esgotar a matriz de taxas terminando numa árvore com o melhor score possível. Por sua vez, o método heurístico apresenta semelhanças ao branch-and-bound, porém ambos diferem no processo de selecção de taxa a adicionar, o método heurístico usa apenas a árvore com o melhor score em cada etapa ao longo do processo. (P. Ajawatanawong, 2017)

2.1.3.3 Método Maximum Likelihood

A *Maximum Likelihood* (ML), ou máxima verossimilhança, é um método estatístico que procura obter o modelo mais provável de ter gerado os dados obtidos através do cálculo de várias versossimilhanças, neste caso estimando parâmetros de um modelo a partir dos dados (sequências de DNA, proteínas entre outros). Este modelo visa estimar as *branch lenghts* e a topologia da árvore com base em modelos de substituição (anteriormente referidos) e alinhamentos. O resultante da análise é uma probabilidade e um modelo contendo a informação genética das matrizes. Este procedimento passa por várias iterações cobrindo todas as possibilidades topológicas a partir de n taxa. Desde que sítios nucleotídicos evoluem independentemente, a árvore é calculada separadamente para cada sítio.

Posto um conjunto de dados e um modelo estatístico, o método de máxima verossimilhança estima os valores dos diferentes parâmetros do modelo estatístico de maneira a maximizar a probabilidade dos dados observados (isto é, busca parâmetros que maximizem a função de verossimilhança). O método de máxima verossimilhança apresenta-se como um método geral para estimação de parâmetros, principalmente no caso de distribuições normais. Para inferir a evolução das sequências, são usados modelos de substituição, dentro destes temos por exemplo o modelo JC64 que funciona assumindo que todas as substiuições ocorrem ao mesmo ritmo(A. Som, 2006) tornando este num dos mais simples destes métodos, acima deste temos o modelo K80 que ao invés de JC64 assume que transições e transversões são eventos diferentes, logo terão diferentes probabilidades de ocorrência(M. Kimura, 1980). Ambos JC64 e K80 tendem à aproximação de um equilibrio, para situações onde é requerido o uso de modelos que apresentam a capacidade de adaptação a disturbios e variações observadas temos modelos como *HKY85* (M. Hasegawa et al., 1985), F81 (J. Felsenstein, 1981), *TN93* (K. Tamura e M. Nei, 1993), *GTR* (L. Gatto et al., 2007) e ademais. De modo a evitarmos erros, como por exemplo, de clustering, é necessário saber usar o modelo que mais se adequa à situação e para tal podemos usar programa que nos indica essa mesma informação, o *ModelTest* (D. Posada e K. A. Crandall, 1998) e o *jModelTest 2* (D. Darriba et al., 2012).

2.1.3.4 Método Maximum Parsimony

Dentro dos métodos de caractéres o método *Maximum Parsimony* (MP) [16, 19], um dos primeiros modelos de inferência a surgir, visa inferir as relações evolutivas de árvores filogenéticas minimizando o número de passos evolutivos necessários à explicação da existência do dados do dataset nas *leaves* (folhas da árvore filogenética). É o método mais usado para a inferência de árvores em que o *dataset* usado apresenta por base um carácter morfológico, na qual é difícil de calcular a taxa (*rate*) de mutação evolutiva. Quando o método MP é aplicado, cada coluna da MSA (*multiple sequence alignment*) é processada como um caracter individual. Porém, nem todas as posições dos alinhamentos são propícias a esta metodologia, como é o caso dos sítios ou locais invariáveis *(invariable sites*). Caracteres (colunas da MSA) que possuem mais do que uma diferença entre os seus nucleótidos ou aminoácidos são designados de *parsimony informative sites*. Apenas estes são usados na inferência pelo método MP, a qual se baseia na pesquisa pela árvore mais parcimoniosa, isto é, aquela que menos passos necessitou para ser construída. Deste modo, a árvore mais curta possível e que consegue inferir relações entre taxa é considerada a mais adequada. Quanto menor for a homoplasia nas sequências, mais preciso será o resultado da inferência (a MP visa minimizar a homoplasia dos alinhamentos). A MP é um critério de otimização simples e intuitivo, podendo ser facilmente aplicado a qualquer tipo de dados, como os *indels.* Porém, é pouco eficiente ao lidar com alinhamentos com elevado nível de variação e matrizes concatenadas com múltiplos genes. (P. Ajawatanawong, 2017)

2.1.3.5 Método Bayesiano

A inferência Bayesiana é um método recente, implementado pela primeira vez na filogenia há cerca de duas décadas atrás (Z. Yang e B. Rannala, 2006). O algoritmo do método Bayesiano infere a árvore com a maior probabilidade posterior (*pp*) (M. E. Alfaro e M. T. Holder, 2006), tendo em conta um grande número de possibilidades, através do algoritmo *Markov chain Monte Carlo* (MCMC) (M. Holder e P. O. Lewis, 2003), aprofundado nas próximas secções. Existem alguns célebres programas filogenéticos que implementam o algoritmo de inferência Bayesiana, como é o caso do *phyloBayes* (A. M. Kozlov, 2019), *MrBayes* [37, 38] e *BEAST2* (R. Bouckaert et al., 2014).

2.1.4 Multispecies Coalescence (MSC)

O modelo MSC descreve a evolução de genes entre espécies, em geral assume que os alelos analisados de cada gene evoluíram de acordo com um processo coalescente em comum dentro de cada espécie. Tipicamente representado como um recuar no tempo começando na ponta de cada ramo na árvore de espécies finalizando na raiz. Logo este modelo MSC modela árvores dentro de árvores e a função densidade $P(T|\theta)$ torna-se mais complexa.
Uma propriedade emergente do modelo MSC conhecido como “Incomplete Lineage Sorting” (ILS) ocorre quando duas ou mais linhagens não coalescem na sua população ancestral imediata, o que pode gerar em árvores de genes com topologias discordantes entre si e com a árvore de espécies. A probabilidade de ocorrência de ILS aumenta à medida que o comprimento dos ramos é reduzido e/ou quando a população efectiva *Ne* é elevada. Árvores de espécies com quatro ou mais espécies no ingroup podem apresentar uma região – zona de anomalia – onde a maior parte das árvores de genes mostram uma topologia diferente à apresentada na árvore de espécies.

Discordância entre árvores de genes e árvores de espécies nas suas topologias e tempos pode levar

a estimativas de árvore de espécie incorrectas em sequências de genes concatenadas – tal tem sido documentado a acontecer em ambos os métodos de máxima verosimilhança e métodos bayesianos (como os que estão implementados em BEAST). Mais especificamente, na zona de anomalia, discordância topológica na árvore de genes pode resultar em estimações incorrectas na topologia da árvore de espécies assim como um bias sistemático nas estimações do comprimento dos ramos. Mesmo no caso onde discordância na árvore de genes é impossível com apenas duas espécies, os tempos de especiação estimados com o método de concatenação serão incorrectos tendo em conta que o tempo de coalescência esperado é 2*Ne* gerações mais antigo que o tempo de especiação, logo a estimação de tempos de especiação com o uso de concatenação serão em previsto, 2*Ne* mais antigos que a verdade [Degnan, James & Rosenberg, 2009].

Ao invés de concatenação, métodos MSC multilocus podem estimar com precisão a topologia e tempos da árvore de espécies e da árvore de genes directamente a partir de múltiplos alinhamentos de sequências (MSAs). A implementação multilocus MSC em BEAST usada neste projecto é denominada como *BEAST, introduzido na versão 1.5.1 de BEAST. Seja $P(T, G, \theta|D)$ a "joint posterior probability density" para a árvore de espécies (T), um set de árvores de genes ($G = \{g1, g2, \ldots, gL\}$) e parâmetros evolucionários adicionais (θ), dados um set correspondente de alinhamentos de múltiplas sequências $D = \{d1, d2, \ldots, dL\}$.

Em seguida, enriquecemos a nossa probabilidade posterior mencionada acima, $P(T, \theta|D)$, através da amostragem adicional de árvores de genes G, usando $P(T, G, \theta|D)$. No MCMC calculamos o producto de verossimilhanças filogenéticas $P(Di|gi, \theta)$, a função de densidade de coalescência $P(gi|T, \theta)$ para cada árvore de genes gi, e a probabilidade prior da árvore de espécies dados parâmetros macroevolucionários $P(T|\theta)$:

$$P(T,G,\theta|D) \propto (\prod_i P(Di|gi,\theta) P(gi|T,\theta)) P(T|\theta) P(\theta).$$

Com a técnica usada na criação de árvores filogenéticas veio-se a verificar que para análises com amostras provenientes dos mesmos indivíduos e referentes às mesmas secções de DNA obtinham-se por vezes árvores filogenéticas completamente diferentes, contrariando o estudo dessas espécies assim como por vezes estas árvores filogenéticas apresentavam anomalias na sua representação com o uso do método de concatenação como por exemplo a existência de tempos de divergência de nódulos mais recentes mais antigos que o tempo de divergência de nódulos ancestrais, transferências horizontais de genes, resultados que proporcionam árvores de genes incongruentes ou seja, embora a árvore resultante da análise possa aparentar ser de boa qualidade com valores de suporte próximos dos 100%, após repetição da análise é normal a árvore resultante ser completamente diferente. Este método de coalescência multiespécies visa mitigar este erro usando árvores das espécies como barreiras de especiação de modo a bloquear a passagem de genes das árvores de genes após divergência impossibilitando certas recombinações e hibridizações.
O programa *BEAST (ou Species Tree Ancestral Reconstruction BEAST) é uma vertente do programa BEAST2 que usa este modelo MSC, deixando o utilizador manipular no separador "Multi Species Coalescent" os parâmetros usados no métodos em que consiste o tamanho da população e a ploidia do dataset em questão, assim como na escolha da árvore de espécies na adição de priors na análise.

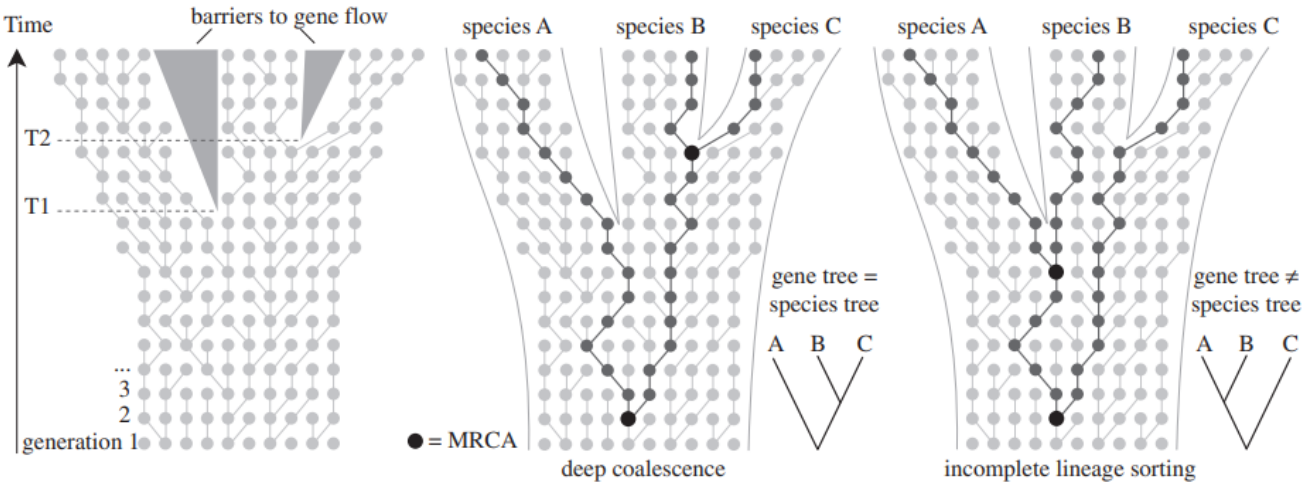


Figura 2 - Coalescence ; Fonte (P. Ajawatanawong, 2017). Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European journal of phycology, 49(2), 179-196.

Uma vertente em que claramente se pode verificar que os métodos coalescentes obtém resultados melhores que os de concatenação é na zona de anomalia, assim como regiões da árvore fora da dita zona porém ainda exibem ILS substancial. A zona de anomalia é caracterizada pela presença de topologias de árvores de genes que apresentam maior probabilidade que a verdadeira árvore de espécies (árvores de genes anômalos = AGTs), que são o resultado inevitável de uma especiação rápida e consecutiva. O método de concatenação não tem hipótese de escapar a zona de anomalia, e irá sempre escolher a árvore errada com o valor de suporte mais elevado. Métodos de coalescência acomodam ILS entre genes em vez de ignorar este processo, logo proporcionam uma solução à zona de anomalia.

2.1.5 Avaliação dos valores de suporte das árvores filogenéticas - confiança

No que diz respeito a métodos filogenéticos uma das maiores fraquezas é o facto dos resultados serem estimativas pontuais da filogenia. De modo que, é necessário dar resposta às perguntas “Quão suportadas são as árvores resultantes do processo de inferência? Como garantimos a sua robustez?”. Tradicionalmente estas questões são encaradas pelo algoritmo estatístico designado de *bootstrap*. A ideia deste método é a matriz original ser aleatoriamente re-amostrada, com substituição dos *sites*, para produzir data sets pseudo-replicados. Quando usados métodos que têm por base critérios de otimização, é iniciada uma *pesquisa de árvore* (*Figura 2-a*, caixa verde) para cada pseudo *dataset*, com base em algoritmos desenvolvidos para tal efeito: *Procura Heuristica* ou *Procura Exacta* (*Tabela 1*). Uma árvore inicial é escolhida de forma aleatória ou a partir do resultado de um

algoritmo, como *Neighbour joining*, *Stepwise addiction* ou *Star decomposition* (*Tabela 1*). A nova árvore é classificada e, se aceite, adicionada à coleção de árvores final. O *bootstrap* é um processo cíclico com termino de acordo com o número de iterações definidas (*Figura 2-a*). O número de vezes que um grupo de sequências ocorre na árvore, durante o processo de amostragem, pode ser usado como medida de quão fortemente o grupo é suportado pelos dados. De forma que, a árvore com melhor *pontuação* de *verosimilhança* será a que terá a que terá maior nível de confiança (M. Holder e P. O. Lewis, 2003).

Em análises Bayesianas, a avaliação dos valores de suporte não tem por base o uso do algoritmo *bootstrap*. Uma vez que, como abordado anteriormente, a estatística Bayesiana assenta na especificação de modelos e *priors* (parâmetros definidos durante o processo de modelação, normalmente assentam sobre valores de distribuição num dado tipo de função) para determinar a probabilidade posterior (*pp*) de cada árvore, tendo em conta a integração dos valores dos parâmetros. De modo que, os algoritmos de *likelihood* se tornam complexos para serem integrados analiticamente em modelos filogenéticos. Assim, os métodos Bayesianos dependem da MCMC para avaliar a confiança das árvores inferidas. A MCMC é um algoritmo notável usado na aproximação de distribuições probabilísticas, numa ampla variedade de contextos. Esta tem por via um processo cíclico (*Figura 2-b*) na qual se cria uma *corrente* com uma serie de passos independentes. A cada passo, uma nova localização para os parâmetros é proposta, de modo a criar uma nova ligação da *corrente*. A localização proposta é similar à usada anteriormente, devido a ser criada aleatoriamente a partir do reajuste dos parâmetros. A densidade da *pp* na nova localização é calculada; caso esta tenha uma *pontuação* superior ao reajuste anterior, é criada uma nova localização e a chain é movida (M. Holder e P. O. Lewis, 2003).

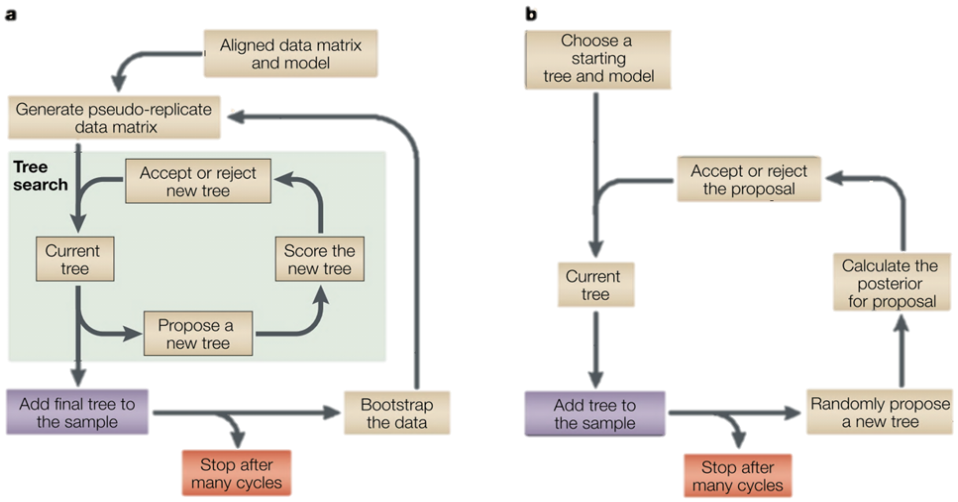


Figura 3 - Workflow de Bootstrap(a) e MCMC(b); (M. Holder e P. O. Lewis, 2003).

2.2 Relógios Moleculares

O "Relógio Molecular" é um termo figurativo para uma técnica que visa usar o ritmo de mutação (mutation rate) de por exemplo sequências de DNA forma a deduzir a altura em que houve divergência, O relógio molecular foi testado pela primeira vez em 1962 numa variante de hemoglobina proveniente de vários animais pelos cientistas Emile Zuckerkandl e Linus Pauling (G.

J. Morgan et al., 1998), que notaram que o número de diferenças na hemoglobina entre linhagens diferentes mudavam linearmente com o passar do tempo, com isto generalizaram esta observação no que depois se veio a chamar a hipótese do relógio molecular referindo que o ritmo de mudanças evolucionárias de qualquer proteina seria aproximadamente constante ao longo do tempo e ao longo de várias linhagens. Esta realização veio a possibilização de estimar a *taxa* de substituições dos aminoácidos, por unidade de tempo, e aplicá-la às diferenças proteicas num grupos de organismos, permitido inferir os tempos de divergências dessas respetivas linhagens. (L. Bromham e D. Penny, 2003)

Com os avanços das tecnologias de sequenciação e o aparecimento de novas ferramentas, veio-se a saber que as taxas de variação genética variavam ao longo das linhagens. O termo relógio molecular refere-se atualmente a um conjunto de métodos e modelos, que inferem a forma de como as taxas de evolução genética variam. Estes métodos possibilitam estimar uma escala de tempo relativa, de modo a permitir obter uma perspetiva cronológica dos grandes acontecimentos que levaram à divergência das espécies, como por exemplo a formação das ilhas Havaianas representada no *Anexo 1*.

2.2.1 Taxa de Variação Evolutiva

Com a informação que as taxas afinal variavam ao longo das linhagens, foram-se categorizando os tipos de taxas, dentro destes temos os site effects que variam com partes do genoma que por exemplo dentro do mesmo genoma temos observações de ritmos evolutivos distintos (Figura 4-A), os efeitos de época que variam ao longo do tempo e por fim temos os lineage effects que variam ao nivel dos taxa. Os *site effects* referidos acima foram os primeiros efeitos a serem caracterizadas e catalogados, durante a investigação genética, sobre taxa heterogéneos. (L. Bromham e D. Penny, 2003)

Os efeitos de linhagem ocorrem quando diferentes taxa apresentam diferentes taxas de evolução molecular (*Figura 4-B*). Os insectos são um óptimo exemplo tendo em conta que a classe insecta apresenta populações com uma elevadíssima taxa de variação genética, parte devido ao curto tempo de geração e parte à quantidade enorme descendentes que proporcionaram mais hipóteses de variação genética. O estudo dos efeitos de linhagem deu origem à metodologia dos relógios relaxados, que visam estatisticamente modelar as taxas de variação ao longo dos ramos da árvore evolutiva. Com o uso de relógios biológicos, este método permite inferir uma escala de tempo evolutivo quando as taxas variam ao longo das linhagens. (L. Bromham e D. Penny, 2003)

Os efeitos de época ocorrem quando as taxas evolutivas diferem em períodos de tempo (*Figura 4-C*). Por exemplo, verificou-se que os taxas evolutivos da gripe aumentaram acentuadamente por volta de 1990. De modo que, esta heterogenia temporal torna-se mais difícil de ser detetada e estudada do que os *site effects* ou as efeitos de linhagem. Isto porque ocorre a criação de padrões de divergência genética ao longo dos taxa que são muito semelhantes aos previstos quando os taxa permanecem constantes ao longo do tempo. (L. Bromham e D. Penny, 2003)

Novas abordagens surgem quando dois ou mais tipos de heterogenias interagem entre si. Os *site*

effects e as efeitos de linhagem interagem quando genes diferentes possuem padrões na taxa de variabilidade distintos ao longo da taxa (*Figura 4*-D). Para lidarmos com estes padrões complexos de variação, podemos socorrer-nos dos modelos de relógio particionados, na qual diferentes partes do genoma evoluem de acordo com diferentes tipos de relógios (*pacemakers)*. (L. Bromham e D. Penny, 2003)

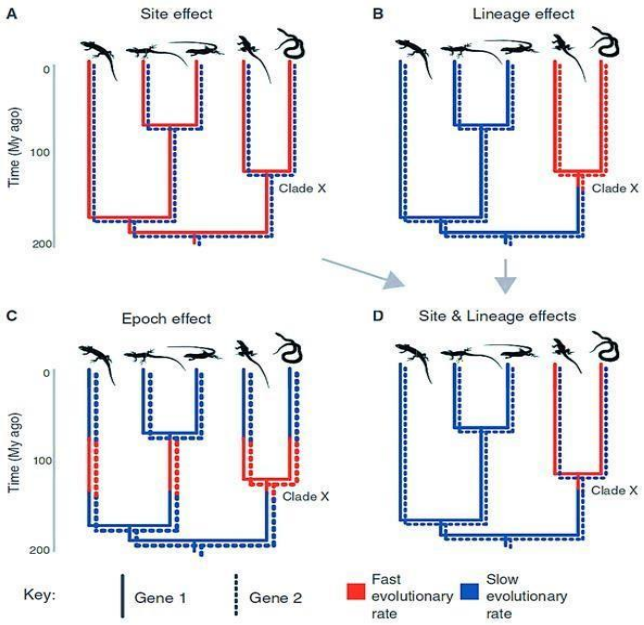


Figura 4 - (A) Taxa de variação ao longo dos sites. (B) Taxa de variação ao longo das linhagens. (C) Taxa de variação ao longo dos períodos de tempo, epochs. (D) Interação dos site effects e efeitos de linhagem*. Fonte (L. Bromham e D. Penny, 2003)*

2.2.2 Escalas de Tempo Evolutivas

Algo que é absolutamente fulcral é o evitar o uso de relógios inapropriados e calibrações incorrectas viso que estas podem originar em resultados falsos e ilusórios podendo distorcer a nossa perspectiva da história evolutiva dos seres vivos enfatizando o uso correcto dos relógios moleculares de modo a manter uma visão correcta da nossa história e não repetir os mesmos erros.

Por ventura, essa situação quase veio a ocorrer com uma análise envolvendo metazoários os quais concluiram que terão divergido há cerca de mil milhões de anos atrás – algo impossível ou no mínimo extremamente improvável pois a datação normal/usual para este tipo de situação tende a estar no intervalo de [542 , 488 Ma] da era Paleozoica. A origem deste erro deve-se à simples falha de se ter considerado os efeitos de linhagem: a variação genética geralmente ocorre mais lentamente em vertebrados do que em invertebrados, mas as primeiras análises moleculares inferiram uma baixa

taxa evolutiva para os vertebrados ao longo do tempo, levando os tempos de divergência animal a estarem associados ao Pré-Câmbrico. (M. S. Y. Lee e S. Y. W. Ho, 2016)

2.2.3 Estatística Bayesiana: Relógios Moleculares

O uso da estatística Bayesiana em filogenética é relativamente recente, mas abre já horizontes no mundo científico, devido ao facto de permitir inferir árvores de incerteza para diferentes grupos taxonómicos presentes na mesma árvore evolutiva. O método Bayesiano está fortemente associado à ML. De modo que, a hipótese ideal é aquela que maximiza a *pp*, que é proporcional à verosimilhança multiplicada pela probabilidade anterior, para cada hipótese. Probabilidades anteriores de diferentes hipóteses possibilitam inferir resultados antes dos dados serem analisados. Em vários métodos, os investigadores definem a distribuição da probabilidade anterior que acreditam ser a mais abrangente, para que a maioria das diferenças na *pp* seja atribuível a diferenças da verosimilhança. Uma das maneiras é aplicar um *prior* uniforme, com a mesma probabilidade a todos os parâmetros possíveis. Para permitir abordagens mais rápidas que a ML *bootstraping*, a inferência Bayesiana permite implementar modelos de sequenciação evolutiva complexos, como a estimativa de tempos de divergência, a deteção de resíduos importantes na seleção natural, bem como pontos de recombinação genética (M. Holder e P. O. Lewis, 2003).

Como já abordado anteriormente, a ML não consegue lidar com modelos que possuem parâmetros muito complexos. Quando a relação entre os parâmetros e os dados é baixa, a inferência pode tornar-se incerta. Na inferência Bayesiana o resultado final é dado tendo em conta todos os parâmetros, ao contrário da ML. Isto porque existem grandes diferenças na forma de como o método Bayesiano e a ML abordam os parâmetros dos modelos evolutivos. Normalmente, a ML tem por base a metodologia *join estimation* que visa encontrar o ponto mais alto do *parameter landscape*. Enquanto que a estatística Bayesiana estima o volume entre um conjunto de *posterior-probabilities*; os parâmetros são integrados entre si, de forma a obter a probabilidade posterior marginal da árvore evolutiva. Além do uso desta metodologia o método Bayesiano usa o algoritmo MCMC (abordado anteriormente no ponto *2.1.3*) para lidar com modelos complexos (M. Holder e P. O. Lewis, 2003).

2.2.3.1 Estatística Bayesiana: Cálculo dos Tempos de Divergência

Com o algoritmo Markov Chain Monte Carlo, é-nos actualmente possível estimar usando estatística Bayesiana os tempos de divergência entre as espécies através da calibração da taxa de substituição, bem como os parâmetros dos modelos evolutivos, mencionados anteriormente: JC, HKY, GTR. Hoje em dia o programa filogenético mais popular que permite, através da modelação da estatística Bayesiana, inferir os tempos de divergência entre as espécies é o *BEAST*, *Bayesian Analysis Sampling Trees* (R. Bouckaert et al., 2014).

O *BEAST* é um programa de inferência Bayesiana, que utiliza o algoritmo MCMC. É totalmente orientado para a inferência dos tempos de divergência de filogenias que têm por base uma *rooted tree* a partir do uso de relógios moleculares *rigoroso* ou *relaxado*. Pode ser usado como um método de reconstrução filogenética, mas também como uma estrutura que visa testar hipóteses evolutivas. O *BEAST* possibilita ao utilizador realizar Inferências de árvores filogenéticas tendo por uso modelos que variam a taxa de substituição (relógios moleculares com *taxa* constante, relógios relaxados não correlacionados, relógios moleculares locais aleatórios); Estimar o tempo de divergência das espécies e calibrações fósseis, através de modelos de tempo de ramo e métodos de calibração; Análisar sequências não-contemporâneas; Aplicar modelos de substituição heterogéneos ao longo das partições de modo a reduzir erros como os de lineage effect visto anteriormente; realizar análises populacionais, como a modelação de parâmetros demográficos (tamanho da população, crescimento/ declínio, migração), criação de *Bayesian skyline plots* e filogeografia o que poderá ser útil no estudo de propagação de vírus; Realização de Inferência a árvores de genes e espécies tendo por base o uso da vertente *Star BEAST*(**BEAST*) (J. Heled e A. J. Drummond, 2010);

Ambos as metodologias, *BEAST* e **BEAST*, estimam a topologia de árvores de espécies, tempos de divergência, tamanho da população entre uma amostra de genes sobre modelos de coalescência multi-espécies, porém existem várias diferenças na modelação. O *BEAST* requere um *outgroup*, o tamanho da população é assumido constante ao longo do ramo e o *prior* para as árvores de espécies é uniforme.

Computacionalmente infere cada gene da árvore individualmente, em duas etapas distintas. Em contraste, o **BEAST* infere a árvore de espécies (métodos *multi-individual*, *multi-locus)*, bem como todos os genes num único processo com o uso do algoritmo MCMC e não necessita de um outgroup (M. S. Y. Lee e S. Y. W. Ho, 2016).

Para este Projecto vai ser usado o *BEAST v2.6.3* (*BEAST2*). Ésta versão é completamente rescrita do programa *BEAST sendo esta mais modular* com a implementação de plugins e packages que proporcionam novas funcionalidades à plataforma como novos modelos de distribuição e por exemplo a SNAPP (*phylogenetic analysis using SNP and AFLP data*) e BDSSM (*birth-death skyline model for serially-sampled data*). A preparação dos ficheiros de *input*, a modelação dos ficheiros de *output*, bem como a visualização é realizada por uma serie de programas disponibilizados pelo *package do* programa *BEAST*, como:

- *BEAUti (Bayesian Evolutionary Analysis Utility)* (R. Bouckaert et al., 2019): um programa com uma *graphical user interface* (GUI) que visa modelar os ficheiros de *input* para o *BEAST* e **BEAST* em formato *eXtensible Markup Language* (XML) assim como é por onde se adicionam novos packages;

- *LogCombiner* (R. Bouckaert et al., 2019) que permite combinar os ficheiros de *log* e de *trees* a partir de múltiplas análises independentes que de certa maneira serve como uma alternativa ao aumento da MCMC chain length proporcionando a opção de combinar resultados de ficheiros xml menores caso o PC usado não seja potente o suficiente;
- *TreeAnnotator* (R. Bouckaert et al., 2019) *que sumariza a informação de um conjunto de árvores numa só*;

- *Tracer* (A. Rambaut et al., 2018) que permite analisar e visualizar a MCMC descrita pelo ficheiro

de log deixando ao utilizador averiguar se os níveis de ESS são adequados;

- *Figtree* (A. Rambaut, 2009) que permite visualizar, sumarizar e anotar árvores filogenéticas. (J. Heled e A. J. Drummond, 2010);

Vários fatores podem influenciar a *taxa* de substituição numa população, tais como a *taxa* de mutação, o tamanho da população, o tempo de geração e a seleção. Como resultado, vários modelos foram desenvolvidos para dar resposta de como a *taxa* de substituição varia ao longo da árvore da vida. Muitos destes modelos são aplicados através de *priors* que têm por base os métodos de inferência Bayesiana.

As implementações dos métodos de datação fornecem uma forma flexível de modelar a *taxa* de variação e obter tempos de divergência com confiança, conferindo se os modelos são adequados. Quando usados com métodos numéricos como a MCMC, para a aproximação da distribuição de parâmetros da *pp*, os métodos Bayesianos demonstram ser extremamente poderosos na inferência dos parâmetros dos modelos estatísticos (R. Bouckaert et al., 2014); o *Anexo 2* demonstra um *workflow* de um processo MCMC.

Várias componentes são aplicadas durante o cálculo dos tempos de divergência com uso dos métodos Bayesianos. Uma delas é o *prior* que define a datação dos nodos, também chamado de *tree prior*. Este descreve como os eventos que deram origem à evolução das espécies estão distribuídos ao longo do tempo. Quando este modelo é combinado com o modelo que calcula a taxa dos ramos, a inferência Bayesiana permite estimar tempos de divergência relativos (R. Bouckaert et al., 2014). No *BEAST*, os *priors* disponíveis para este tipo de abordagem, cálculo dos tempos de divergência inter-espécies, são variantes do *birth-death prior*, que incluem o *calibrated Yule model*, o modelo *birth-death* com amostras incompletas das espécies, bem como *serially- sampled birth-death processes* (R. Bouckaert et al., 2014).

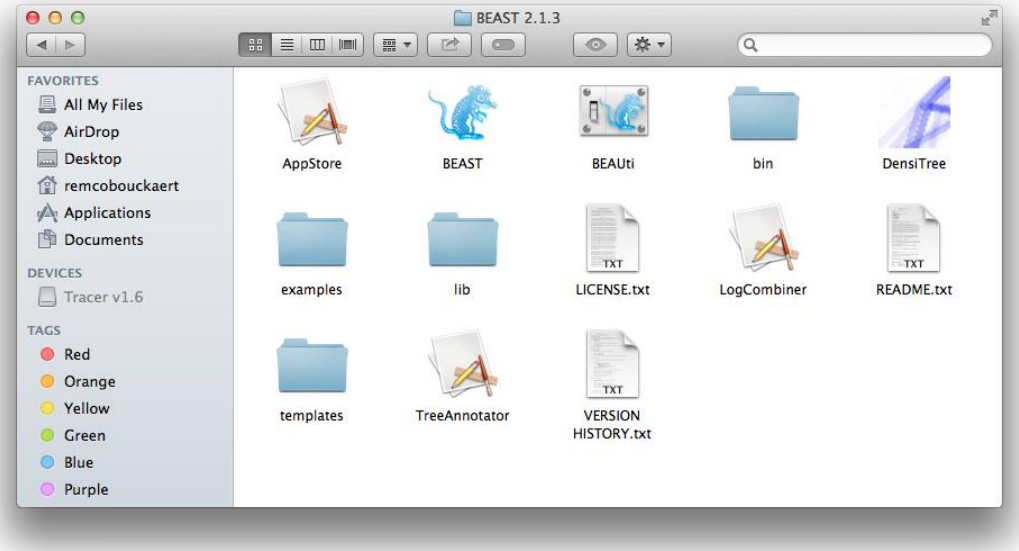


Figura 5 - Exemplo de pasta com alguns dos programas *mencionados do framework BEAST*

3 Metodologia

As atividades desenvolvidas ao longo do projeto tiveram todas como ponto de partida a seleção das sequências a serem analisadas, a criação de datasets com as amostras, bem como a análise da qualidade das sequências, manipulação e estruturação nucleotídica após o processo de alinhamento.

As sequências estudadas dos escaravelhos tigre (Coleoptera Cincidelini) e dos escaravelhos do grupo *Trechus fulvus* foram adquiridas a partir do portal *National Center for Biotechnology Information* (NCBI) com base na API (*a*pplication programming interface*) -eutils* (que simplifica a pesquisa, aquisição, e análise de registos do NCBI) do sistema *Entrez* (*Global Query Cross-Database Search System*) desenvolvido pela plataforma da NCBI: o comando *efetch* (usado para fazer o *download* de amostras a partir da base de dados NCBI) com base nos números de acesso facultados pelas respectivas publicações. O *script* foi executado na *aplicação GitBash* do sistema operativo *Windows 10*, criando um ficheiro em formato *FASTA* na diretoria desejada com todas as sequências selecionadas;

As sequências em formato *FASTA* são em primeiro lugar alinhadas com uso do programa *MAFFT v7.419* (K. Katoh e D. M. Standley, 2013), com os parâmetros *standard (--auto)*. Após o alinhamento, o *dataset* é analisado com o programa *AliView V1.26* (A. Larsson, 2014) de forma a inferir a qualidade dos alinhamentos. Em alguns casos foi necessário cortar (*trimming*) as extremidade das sequências para que tivessem igual número de pares de bases.

Por último o ficheiro *FASTA* é convertido em formato *Nexus* para ser usado como *input* nos seguintes passos. O primeiro passo será com recurso ao programa *jModelTest2*, estimar para cada *dataset*, qual o modelo de substituição que melhor se adequa a cada *dataset*.

3.1 Estudo do enquadramento do programa *BEAST* na inferência filogenética molecular

Como ponto de partida do projeto, foram efetuadas as leituras dos documentos de apoio disponibilizadas *online* pelo programa *BEAST*, quer da versão 1 e 2 (nomeadamente o "Taming the BEAST") para inferência *inter* e *intraespecífica*, bem como a realização de tutoriais. Os materiais estudados de maior relevância, estão apresentados no *Anexo 3*.

Esta primeira fase foi fundamental para interiorizar as bases necessárias ao uso e enquadramento do programa na inferência filogenética, tendo por via os relógios moleculares, modelos de substituição e escolha de priors, ou seja, todos os passos envolvidos na análise. Foram ainda pesquisadas e estudadas as publicações mais adequadas ao projeto, apresentadas no *Anexo 4* [54-59].

3.2 Reprodução do paper referente à inferência molecular dos escaravelhos tigre (Coleoptera: Cicindelini), e descrição da pipeline comum do programa BEAST

Neste ponto, foi reproduzida a análise da publicação "*Phylogeographic patterns of two tiger beetle species at both sides of the strait of Gibraltar (Coleoptera: Cicindelini)*" (García-Reina,A. et al., 2014).

O artigo retrata a inferência interspecífica com recurso à metodologia *BEAST2*, contudo foi realizada uma análise intraespecífica adicional, de modo a comparar os resultados das duas abordagens. Para a inferência filogenética molecular com o programa *BEAST2*, as sequências em formato Fasta são em primeiro lugar alinhadas e formatadas em ficheiros *Nexus* segundo a metodologia usada neste projecto (desenvolvida na próxima secção) foram carregadas no programa *BEAUti* de modo a calibrar os parâmetros da MCMC e criar o ficheiro de controlo (ficheiro em formato XML que irá ser interpretado pelo programa). Os parâmetros foram aplicados com base nas informações dadas pela publicação (*Tabela 2)*. Após a criação do ficheiro XML, foi executado o programa *BEAST2* e carregado o ficheiro. Quando concluído o processo da MCMC, obteve-se um ficheiro de *log* e um ficheiro em formato *trees*.

Parâmetros / Genes	Site Models	Clock Model	Priors	MCMC
COI	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.08606	Coalescent Constant Population (Parâmetros Default) Cephalota Outgroup (Monophyletic Prior)	10M de iterações guardadas a cada 100;

Tabela 2 - Parâmetros usados na calibração do gene mitocondrial COI das espécies de escaravelhos tigre, formatados para BEAST2.

Para a inferência intraespecífica, **BEAST2*, foi criado um novo ficheiro XML com a calibração apresentada na *Tabela 3, no* programa *BEAUti*. Apesar de uma diferente formatação o ficheiro de controlo da é igualmente processado pela metodologia *BEAST2*.

Parâmetros / Genes	Site Models	Clock Model	Multi Species Coalescent	Priors	MCMC
COI	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.08606	Species Tree Population Size: Pop Function = Linear with Constant Root Population Mean = 1.0 Ploidy = Y or Mitochondrial	Coalescent Constant Population (Parâmetros Default)	10M de iterações guardadas a cada 100;

*Tabela 3 - Parâmetros usados na calibração do gene mitocondrial COI das espécies de escaravelhos tigre (Coleoptera: Cicindelini), formatados para *BEAST2.*

Após o processamento do ficheiro xml na aplicação, foi carregado o ficheiro de *log* no programa *Tracer* (de modo a averiguar a qualidade dos resultados obtidos) e verificados os valores de ESS (*Effective Sample Size*) dos parâmetros (exemplo apresentado nos *Anexos 5 e 6*), onde são testadas as *runs*/"leituras do ficheiro xml" independentes da pp (posterior probability). Caso os parâmetros convergirem correctamente (valores acima de 200 para melhor efeito, embora entre 100 e 200 ainda seja viável), eram efetuadas mais uma série de *runs* independentes (com *seeds* diferentes, definição default), e consequentemente os ficheiros de *log* eram combinados a partir do programa LogCombiner, bem como os ficheiros em formato trees também no mesmo programa; caso contrário, os parâmetros seriam novamente calibrados *no BEAUti* (caso os valores de ESS não sejam superiores a 200 após várias tentativas dá-se nesse caso como o seu próprio resultado útil para a comparação entre os dois métodos).

Devido à inferência ser suportada por métodos estocásticos, por vezes a distribuição acaba por não convergir corretamente. A combinação de *logs* permite garantir um valor de suporte às retas de distribuição dos *priors* ao longo das gerações do MCMC, levando a completarem-se entre si. Neste caso, foram combinados 8 ficheiros de *log* e de *trees*, a partir de 8 *runs* independentes, após várias calibrações. Seguidamente, os ficheiros de *trees* foram combinados com 10% de *burn-in* e anotados com o programa *TreeAnnotator*, de modo a reportar a *Maximum Clade Credibility Tree*, com os nodos calibrados pela altura média. De modo a não se perder informação, não foi aplicado novo *burn-in* à anotação devido ao facto de este já ter sido aplicado anteriormente na combinação. Quando terminado o processo de anotação, com uma grande necessidade de tempo de processamento (aproximadamente 20 a 30 minutos dependendo da memória livre do PC), o *output* (a árvore filogenética final) foi analisado e editado a partir do programa *FigTree*.

Esta metodologia retrata o *workflow* comum à inferência filogenética molecular com as metodologias *BEAST2* e **BEAST2* usado neste projecto (*Figura 4*). As seguintes análises partilharam do mesmo método retratado. A análise de resultados desta inferência não faz parte dos objetivos do projeto, uma vez que este apenas serve como *dataset* de treino por abordar a vertente **BEAST2*.

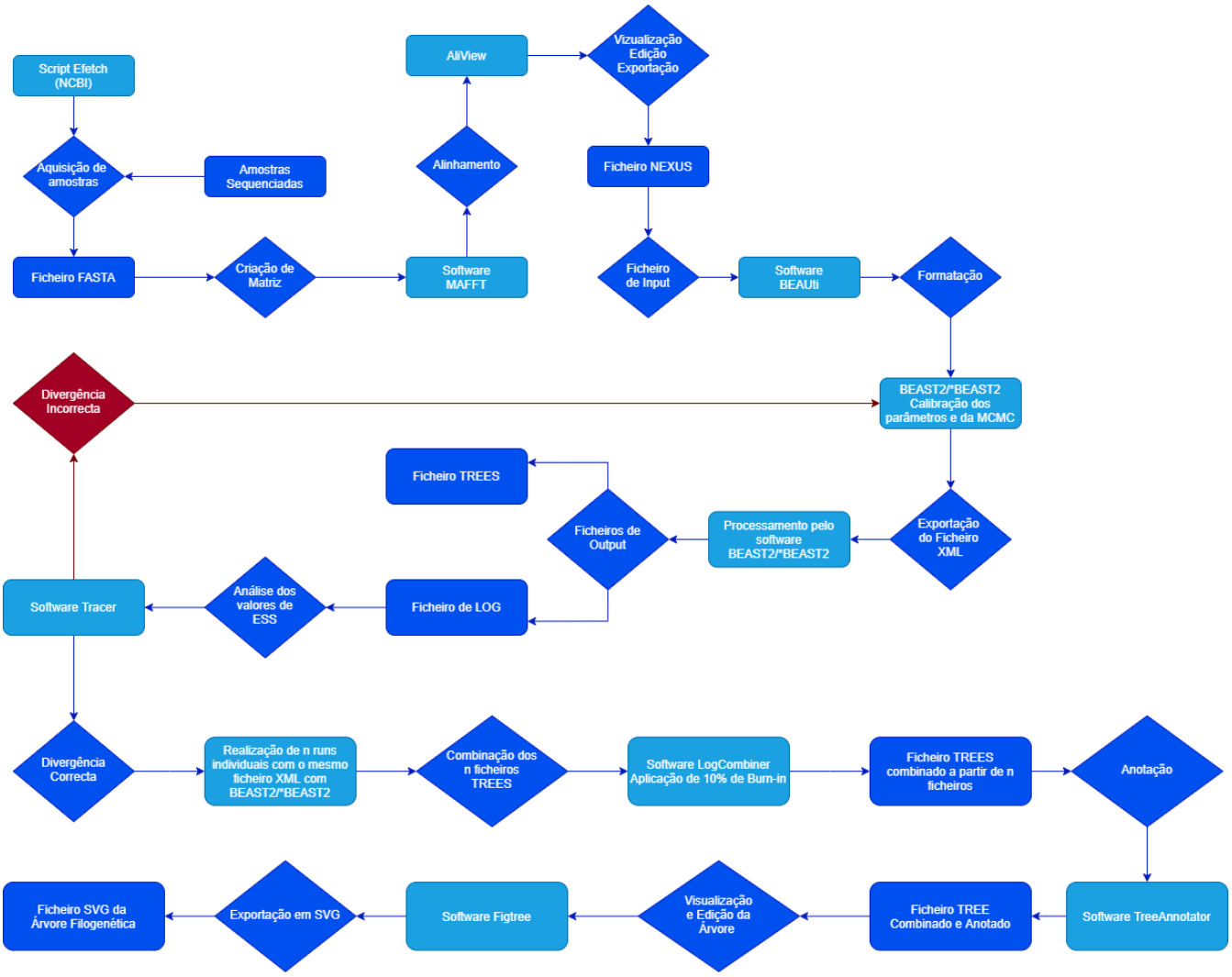


Figura 6 - Workflow usado no projecto

3.3 Reprodução do paper relativo à inferência filogenética molecular da espécie *Trechus fulvus*

Neste terceiro ponto foram replicados os resultados da publicação "Late Miocene origin of an IberoMaghrebian clade of ground beetles with multiple colonizations of the subterranean environment" (A. Faille et al., 2014). Com os *accession numbers* do *GenBank* disponibilizados pela publicação, foram descarregadas as sequências do gene mitocondrial *Cytochrome c Oxidase Subunit I* (*COI)* em formato *FASTA* pelo método *efetch*, anteriormente descrito, e manipuladas com o programa *AliView* após o alinhamento com o programa *MAFFT*, seguidamente exportadas em formato *Nexus*.

Procedeu-se à aplicação da metodologia descrita no workflow.O dataset em formato *Nexus* do *Cytochrome c Oxidase Subunit I* foi carregado no programa *BEAUti* de modo a calibrar os parâmetros da MCMC, e criar o ficheiro XML para ser processado pelo *BEAST2*. Os parâmetros usados na calibração da inferência interespecífica estão designados na *Tabela 4*.

Parâmetros / Genes	Site Models	Clock Model	Priors	MCMC
COI	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.015 (Lower bound = 0.01 Upper bound = 0.0198)	Yule model (Parâmetros Default) Trechus Fulvus Ingroup (Monophyletic Prior)	10M de iterações guardadas a cada 100;

Tabela 4 - Parâmetros usados na calibração do gene mitocondrial Cytochrome c Oxidase Subunit I do grupo , para inferência interespecífica com o método BEAST2.

Quando finalizado o processamento, o ficheiro de log resultante foi analisado no programa Tracer, até encontrarmos os parâmetros adequados, de acordo com os valores da distribuição de ESS. Após várias tentativas de calibração, procedeu-se à execução de novas análises independentes (com diferentes seeds) a partir do mesmo ficheiro XML para serem combinadas num único ficheiro a partir do programa LogCombiner, bem como o ficheiro de trees combinado com a informação das novas runs com 10% de burn-in. Procedeu- se à anotação com o programa TreeAnnotator sem burn-in e os nodos calibrados pela Mean Height, prosseguindo-se a análise com a visualização da árvore com a maior credibilidade no Figtree.

Foi seguidamente feita a análise intraespecífica do mesmo *dataset*, tendo como metodologia o workflow usado no passo anterior, mas formatado no *BEAUti* para **BEAST2*, e calibrado com os parâmetros apresentados na *Tabela 5*.

Parâmetros	Site Models	Clock Model	Multi Species Coalescen	Priors	MCMC
COI	Modelo de substituição JC69 Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.015 (Lower bound = 0.01 Upper bound = 0.0198)	Population Function : Linear with Constant Root Pop mean = 0.35 (Lower bound = 0.3 Upper bound = 0.4) Gene Ploidy = Y or Mytochondrial	Yule model (Parâmetros Default) Trechus Fulvus Ingroup (Monophyletic Prior)	10M de iterações guardadas a cada 100;

*Tabela 5 - Parâmetros usados na calibração do gene mitocondrial Cytochrome c Oxidase Subunidade I das espécies de escaravelhos do grupo , para inferência intraespecífica com o método *BEAST2.*

3.4 Inferência filogenética molecular de escaravelhos (Coleoptera: Cicindelini)

Esta tarefa consistiu em calcular os tempos de divergência para os escaravelhos (Coleoptera: Cicindelini) com apenas a vertente interespecifica do programa *BEAST2*, tendo em conta a sistemática biológica. A informação genética foi obtida através do portal NCBI a partir dos accession numbers usados nas publicações. Cada gene foi calibrado individualmente, e só depois foi realizada a calibração concatenada.

3.4.1 Inferência filogenética molecular de escaravelhos (Coleoptera: Cicindelini) – BEAST2

Primeiramente, procedeu-se à criação dos *datasets*. Neste procedimento foram utilizados dois genes; o gene mitocondrial *Cytochrome c Oxidase subunidade I* (*COI*) e o gene mitocondrial 16S. Após as matrizes devidamente criadas e convertidas no formato *Nexus*, procedeu-se ao uso do workflow. A calibração foi realizada de acordo com os parâmetros apresentados na *Tabela 6*.

Parâmetros / Genes	Site Models	Clock Model	Priors	MCMC
COI	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.08606	Coalescent Constant Population (Parâmetros Default) Cephalota Outgroup (Monophyletic Prior)	10M de iterações guardadas a cada 100;
16S	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.0054		

Tabela 6 - Parâmetros usados na calibração do dataset concatenado composto pelos gene mitocondrial COI e 16S dos escaravelhos (Coleoptera: Cicindelini), formatados para BEAST2.

3.4.2 Inferência filogenética molecular de escaravelhos (Coleoptera: Cicindelini) - *BEAST2

Para a concretização desta tarefa, foi novamente usado como base de operações o workflow. Os parâmetros usados na calibração dos genes estão descritos na *Tabela 7*.

Parâmetros / Genes	Site Models	Clock Model	Multi Species Coalescent	Priors	MCMC
COI	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.08606	Species Tree Population Size: Pop Function = Linear with Constant Root Population Mean = 1.0 Ploidy = Y or Mitochondrial Partições dos genes linked para a Inferência da árvore, de modo a partilharem branch times e a mesma topologia	Coalescent Constant Population (Parâmetros Default)	10M de iterações guardadas a cada 100;
16S	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.0054			

*Tabela 7 - Parâmetros usados na calibração do dataset concatenado composto pelos gene mitocondrial COI e 16S dos escaravelhos (Coleoptera: Cicindelini), formatados para *BEAST2.*

3.5 Inferência de *Trechus fulvus* – dataset concatenado

De modo a complementar a inferência realizada no *ponto 3.3*, realizou-se uma nova inferência para o estudo das escaravelhos do grupo Fulvus. Foi realizada a análise de mais genes mitocondriais, *Cytochrome c Oxidase Subunidade I (COI)* e *rrnL+trnL+nad1* e dois nucleares, *LSU* e *SSU*. À semelhança da abordagem realizada só com o *Cytochrome c Oxidase Subunidade I*, as amostras foram adquiridas com base no *script efetch (Apêndice 1)*, pelos *accession numbers* do *GenBank* providenciados pela publicação. Após analisada a qualidade das sequências com recurso ao programa *AliView* e convertidas para formato *Nexus*, recorreu- se ao *workflow* deste projecto de modo a inferir a *Maximum clade credibility tree* com base nos parâmetros apresentados na *Tabela 8* para a análise interespecífica com o programa *BEAST2*, e a *Tabela 9* para a análise intraespecífica com recurso à dependência **BEAST2*, de modo a estimar os tempos de divergência relativos das espécies de escaravelhos do grupo Fulvus. Os genes foram calibrados individualmente e de forma concatenada.

Parâmetros / Genes	Site Models	Clock Model	Priors	MCMC
COI	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.015 (Lower bound = 0.01 Upper bound = 0.0198)	Yule Model (Parâmetros Default) Trechus Fulvus Ingroup (Monophyletic Prior)	50M de iterações guardadas a cada 1000;
SSU	Modelo de substituição GTR: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Relaxed Clock Log Normal Clock.Rate = 3.0E-4 (Lower bound = 1.0E-4 Upper bound = 4.0E-4)		
LSU	Modelo de substituição GTR: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Relaxed Clock Log Normal Clock.Rate = 0.001025 (Lower bound = 7.0E-4 Upper bound = 0.002)		
rrnL+trnL+nad1	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.016 (Lower bound = 0.01 Upper bound = 0.022)		

Tabela 8 - Parâmetros usados na calibração do dataset concatenado dos escaravelhos do grupo composto pelos genes mitocondriais rrNL+trnL+nad1 e COI assim como genes nucleares SSU e LSU; formatados para BEAST2.

Parâmetros / Genes	Site Models	Clock Model	Multi Species Coalescent	Priors	MCMC
COI	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.015 (Lower bound = 0.01 Upper bound = 0.0198)	Population Function : Linear with Constant Root Pop mean = 0.35 (Lower bound = 0.3 Upper bound = 0.4) Gene Ploidy = Y or Mytochondrial	Yule Model (Parâmetros Default) Trechus Fulvus Ingroup (Monophyletic Prior)	50M de iterações guardadas a cada 1000;
SSU	Modelo de substituição GTR: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Relaxed Clock Log Normal Clock.Rate = 3.0E-4 (Lower bound = 1.0E-4 Upper bound = 4.0E-4)			
LSU	Modelo de substituição GTR: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Relaxed Clock Log Normal Clock.Rate = 0.001025 (Lower bound = 7.0E-4 Upper bound = 0.002)			
rrnL+trnL+nad1	Modelo de substituição JC69: Rate de substituição = 1.0 (não estimado) Gamma Category Count = 0 Proportion Invariant = 0 (não estimado)	Strict Clock Clock.Rate = 0.016 (Lower bound = 0.01 Upper bound = 0.022)			

*.Tabela 9 - Parâmetros usados na calibração do dataset concatenado dos escaravelhos do grupo composto pelos genes mitocondriais rrNL+trnL+nad1 e COI assim como genes nucleares SSU e LSU; formatados para *BEAST2.*

4 Descrição de resultados

4.1 Espécies de escaravelhos tigre (Coleoptera: Cicindelini)

Nas *Figuras 7* e *8* estão apresentadas as árvores filogenéticas dos escaravelhos tigre (Coleoptera: Cicindelini) provenientes da Região de Gibraltar, inferidas através das abordagens *BEAST2* e **BEAST2* respetivamente, resultante da metodologia apresentada no *ponto 3.4* e calibradas com base nos parâmetros da *Tabela 6* e *7*.

As árvores filogenéticas resultantes de ambas as inferências realizadas neste projeto são compostas por três grandes *clades*. Os *Lophyra - Cicindela* (*clade I*), *C.ibérica - Cincidela* (*clade II*) e *C.campestris - C.maroccana* (*clade III)*. Os resultados obtidos pelo método *BEAST2*, sugerem que a separação entre Lophyra e Cicindela deverá ter ocorrido no intervalo [2.2 , 2.65 Ma] com um valor de suporte *pp*=1, a separação de *C.ibérica* e *Cicindela* terá ocorrido no intervalo [1.4 , 1.95 Ma], suportada pela *pp*=0.9999 e linhagem *C.campestris* terá separado de *C.maroccana* no intervalo [0.65 , 1.15 Ma] com *pp*=1.

Com o método **BEAST2* verifica-se que a divergência dos Lophyra e Cicindela poderá ter ocorrido no intervalo [3.66 , 4.58 Ma] com *pp*=0,9907, a separação dos *C.ibérica* e C.campestris no intervalo [2.14, 3.2 Ma], com um valor de suporte *pp*=0.9991 e a linhagem *C.campestris* separa-se de *C.maroccana* no intervalo [0.89 , 2.3 Ma] com pp=0.997

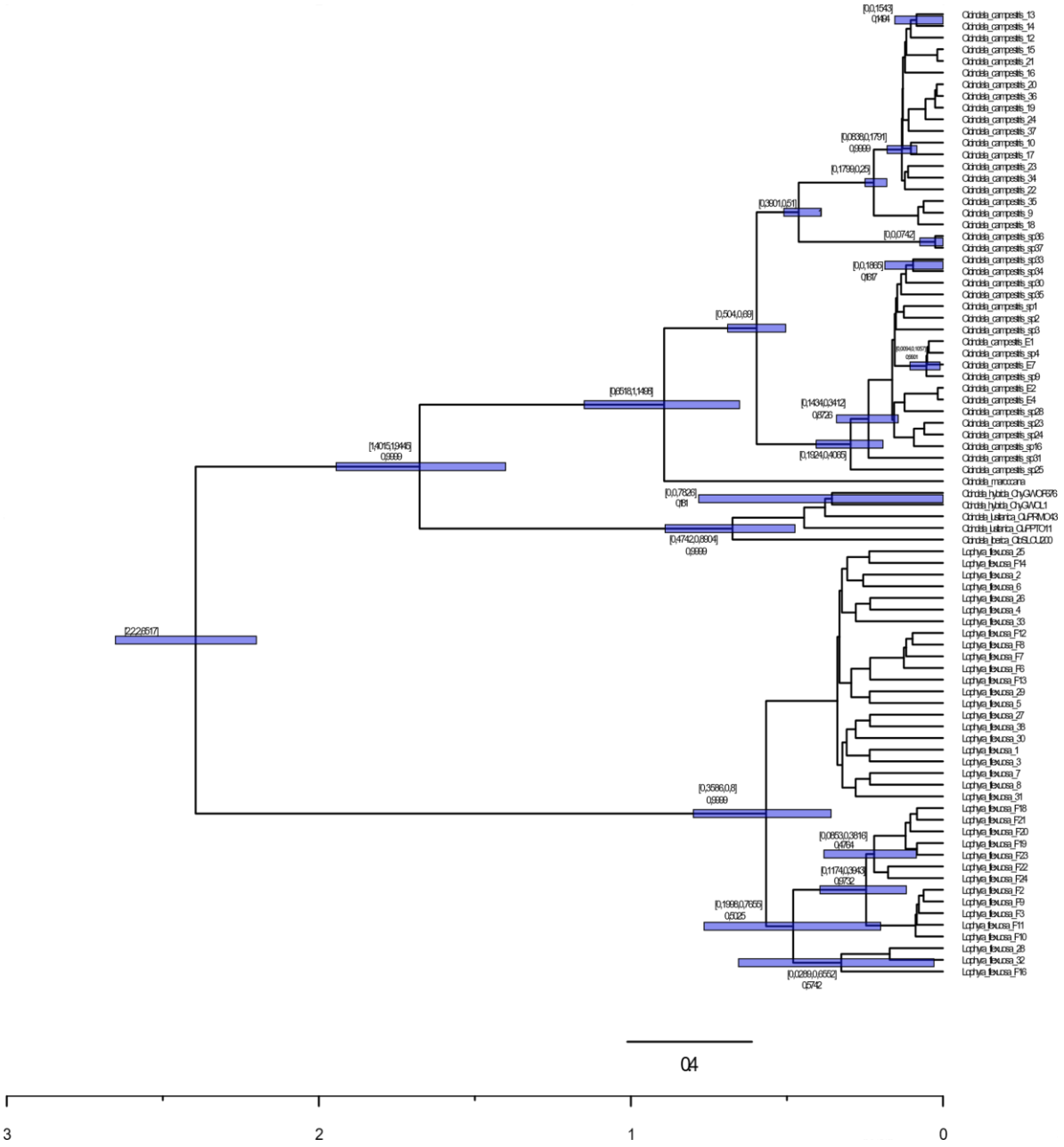


Figura 7 - Árvore filogenética resultante da inferência interespecífica (BEAST2) das espécies de escaravelhos tigre (Coleoptera: Cicindelini), calibrada com os parâmetros da Tabela 2; dataset composto pelo gene mitocondrial COI e 16S, inferido pela combinação de 8 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.

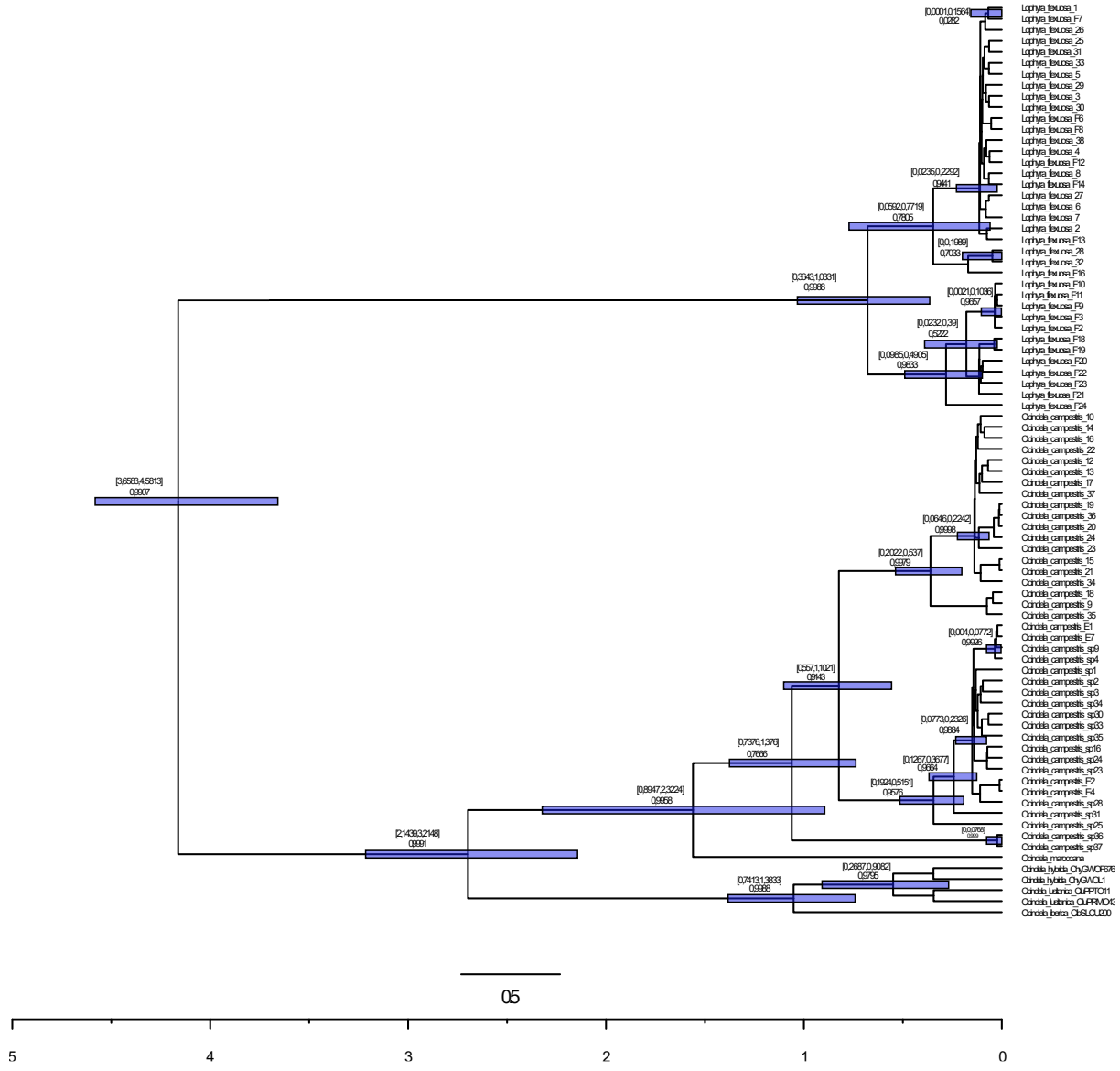


*Figura 8 - Árvore filogenética resultante da inferência intraespecífica (*BEAST2) das espécies de escaravelhos tigre (Coleoptera: Cicindelini), calibrada com os parâmetros da Tabela 3; dataset composto pelo gene mitocondrial COI e 16S, inferido pela combinação de 8 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.*

4.2 Escaravelhos do Grupo Fulvus (*Trechus fulvus*) – dataset constituído pelo gene COI

4.2.1 Inferência interespecífica, BEAST2

Na *Figura 9 (e Figura 10)* está apresentada a árvore filogenética resultante da inferência interespecífica ao gene *COI* das espécies de *Trechus fulvus*, resultante da metodologia descrita no ponto 3.3. Esta foi calibrada de acordo com os parâmetros apresentados na *Tabela 4* e obtida a partir da combinação de 3 *runs* independentes.

Topologicamente, o *outgroup* é constituído por vários espécies distintas, e o *ingroup* por vários individuos da espécie *Trechus fulvus*. Os *clades* que iremos ter em conta para a discussão serão os *Clades A* a *L* com o *clade A* sendo a "raíz" deste grupo Fulvus representado no gráfico como o nódulo que encompassa a espécie *gigoni* à *delhermi*, o *clade B* é representado pelo próximo nódulo com a ausência da espécie *delhermi*, passando ao *clade C* temos a linhagem principal de e a espécie *iblanensis* e de seguida ao *clade D* que apresenta as espécies *gigoni* à *martinezi*. A separação entre a espécie delhermi do restante grupo Fulvus poderá ter ocorrido no intervalo [16.1 , 16.3 Ma] com valor *pp*=1, durante a época Mioceno. Os *clades C* e *D* (*clade B*) poderão ter divergido entre si no intervalo [7.1 , 7.6 Ma] novamente com suporte (*pp*=1). passando para o clade C a linhagem principal de *Trechus fulvus* poderá ter divergido da espécie iblanensis no intervalo [5.2 , 7.16 Ma], com um valor de suporte (*pp*=1).

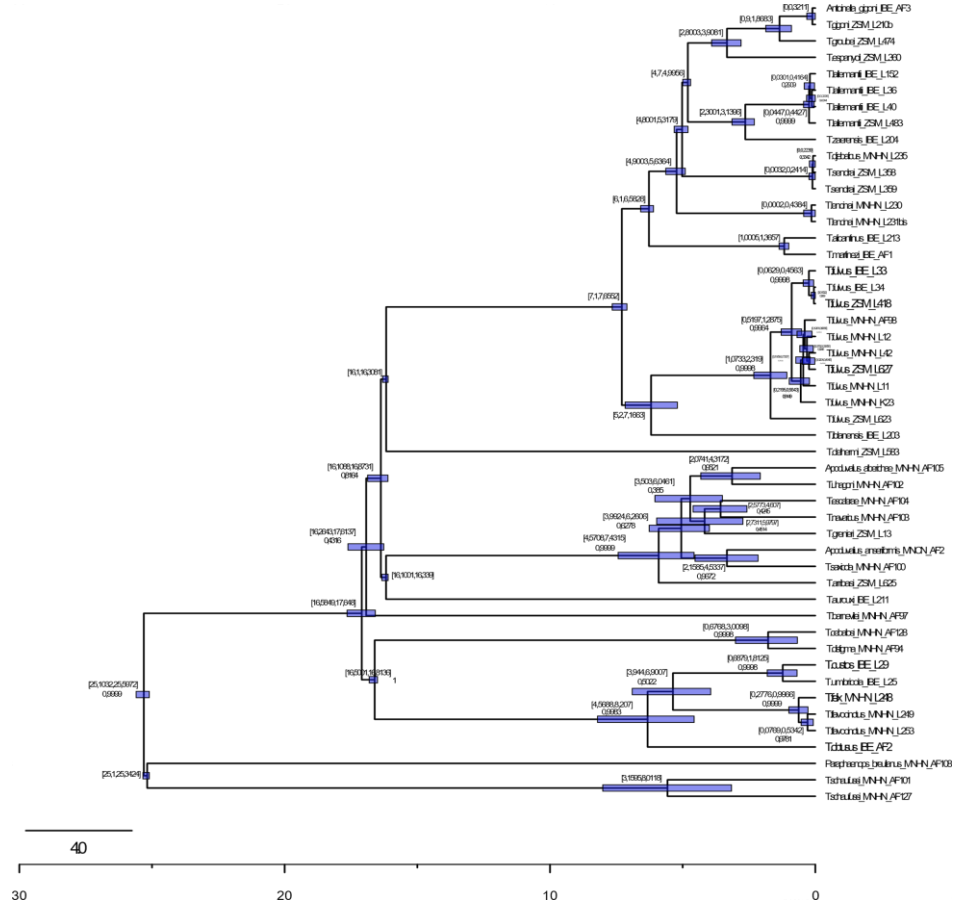


Figura 9 - Árvore filogenética resultante da inferência interespecífica (BEAST2) ao dataset de escaravelhos do grupo Fulvus composto pelo gene mitocondrial COI, calibrado com os parâmetros da Tabela 4 a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.

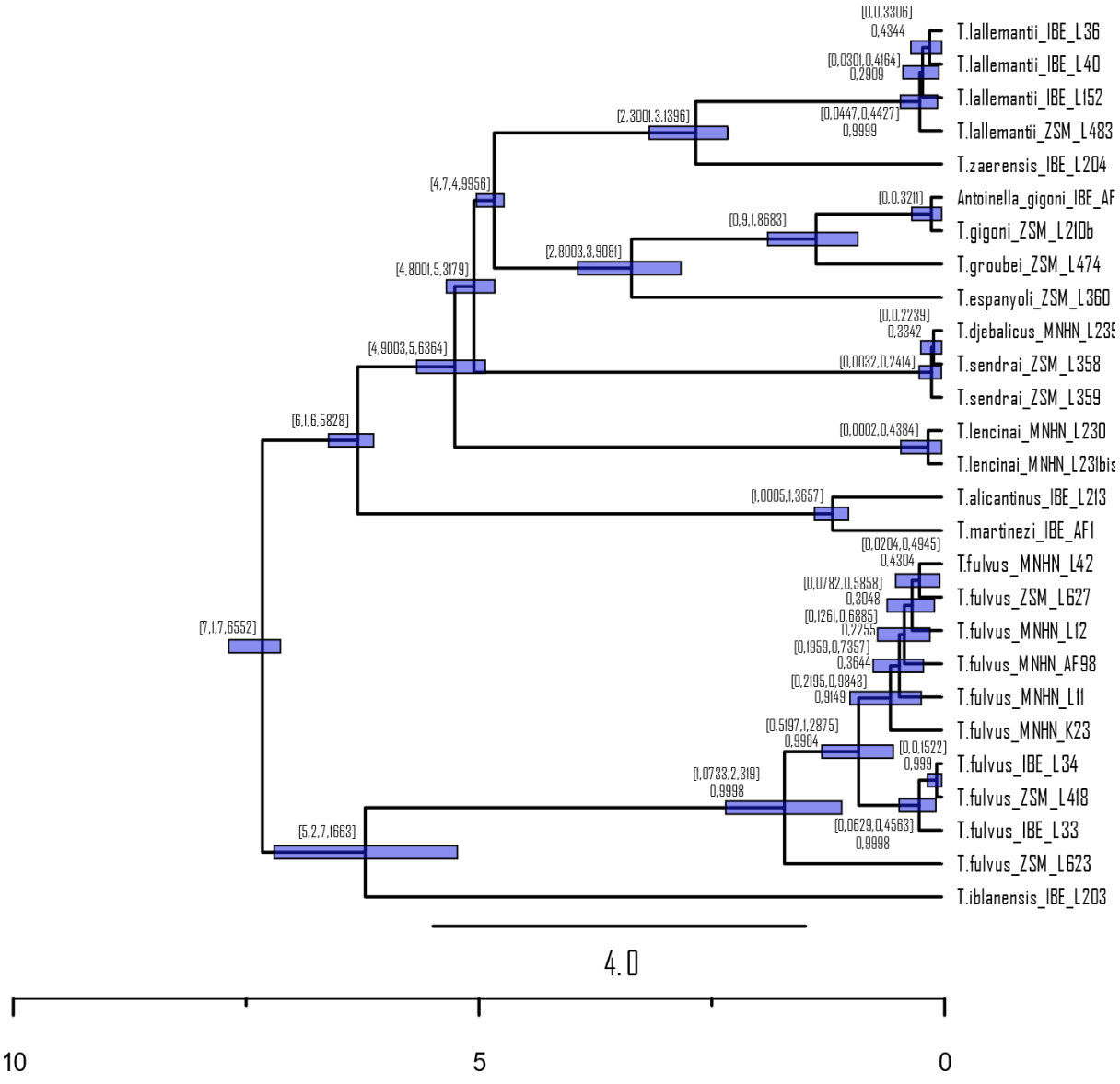


Figura 10 - Ampliação da Árvore filogenética resultante da inferência interespecífica (BEAST2) ao dataset de escaravelhos do grupo Fulvus composto pelo gene mitocondrial COI, calibrado com os parâmetros da Tabela 4 a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.

4.2.2 Inferência intraespecífica, *BEAST2

Continuando com a nomenclatura da inferência anterior vamos considerar os mesmos clades para a figura 11 (e figura 12) onde podemos começar com o intervalo observado no *clade A* onde temos um intervalo entre [16.1 , 16.59 Ma] com um valor relativamente baixo *pp*=0.6423 continuando a suportar o artigo original em que o grupo Fulvus originou na Época Mioceno (A. Faille et al., 2014). Deixando a espécie delhermi para trás observamos que temos um intervalo entre [7.11 , 7.83 Ma] com um suporte elevado (*pp*=1) no *clade B* dividindo assim os *clades C* e *D*, continuando assim para o *clade C* onde temos a separação entre a linhagem principal e a espécie iblanensis onde verificamos um intervalo entre [5.36 , 7.48 Ma] com um valor razoável de suporte (*pp*=0.9272).

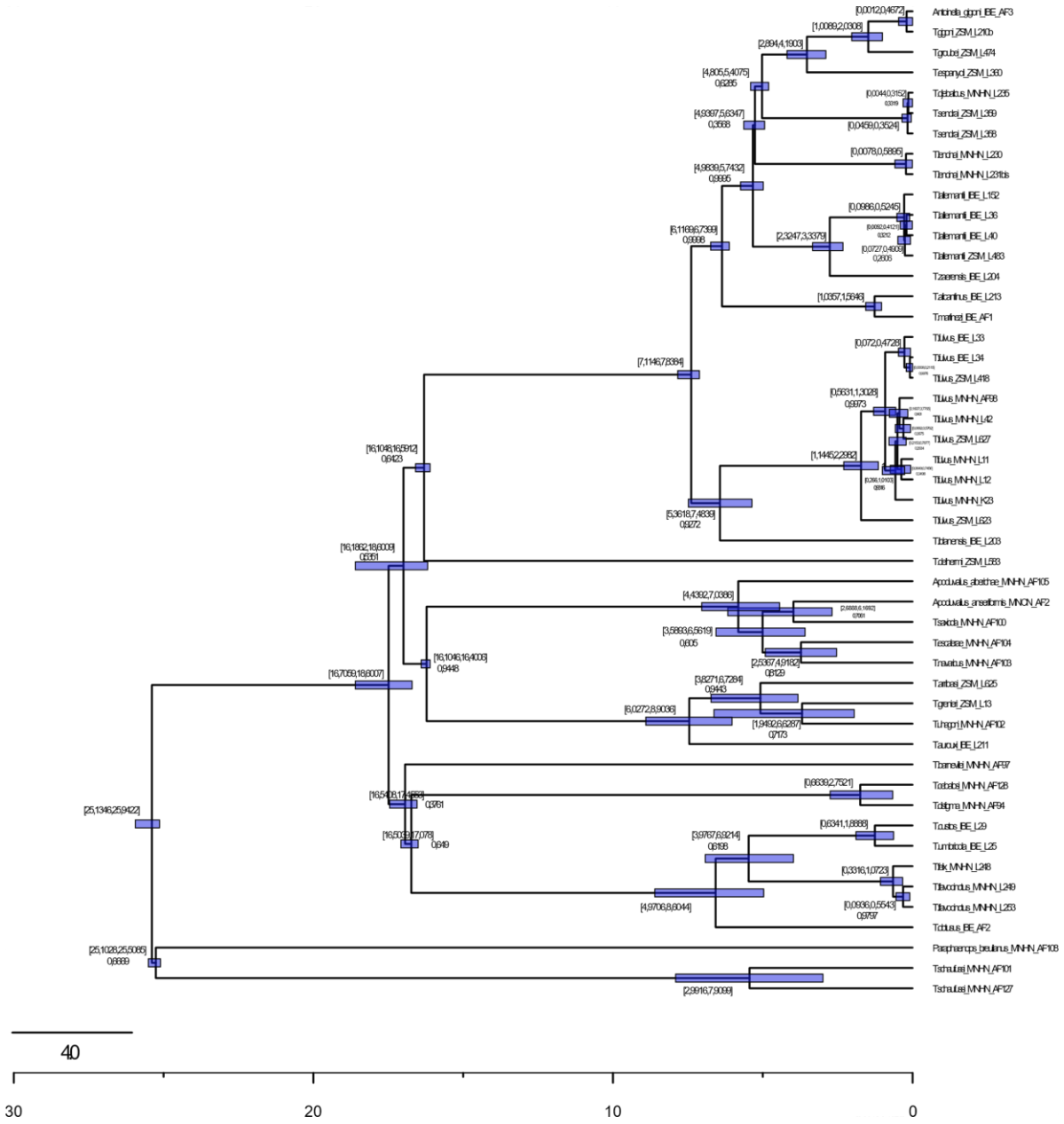


*Figura 11 – Árvore filogenética resultante da inferência intraespecífica (*BEAST2) ao dataset de escaravelhos do grupo Fulvus composto pelo gene mitocondrial COI, calibrado com os parâmetros da Tabela 4 a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.*

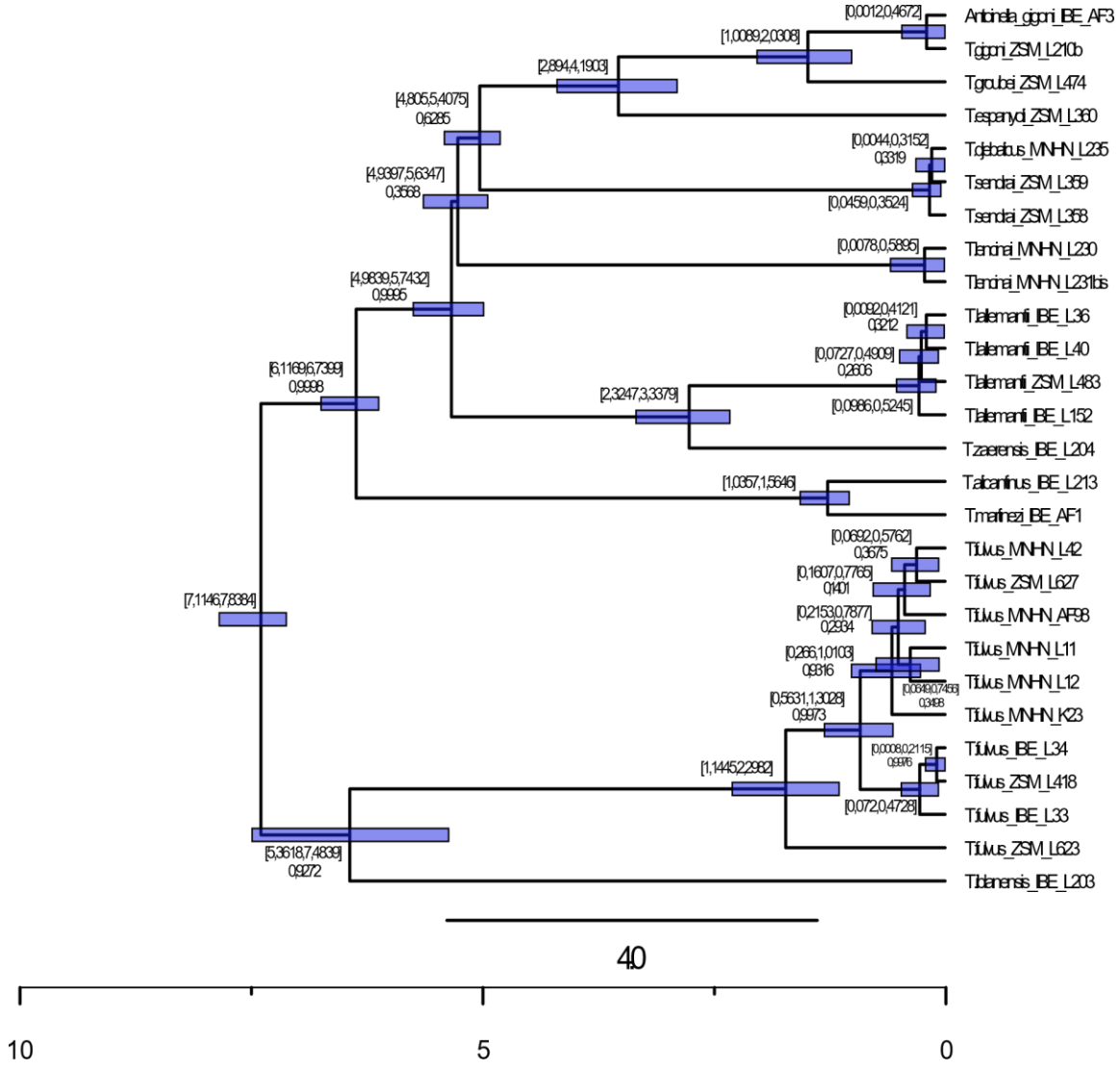


*Figura 12 – Ampliação da Árvore filogenética resultante da inferência intraespecífica (*BEAST2) ao dataset de escaravelhos do grupo Fulvus composto pelo gene mitocondrial COI, calibrado com os parâmetros da Tabela 4 a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.*

4.3 *Trechus fulvus* – dataset concatenado

4.3.1 Inferência interespecífica, BEAST2

A *Figura 13 (e figura 14)* apresenta a árvore filogenética das espécies de escaravelhos do grupo Fulvus (*dataset* constituído pelos genes *COI, SSU, LSU e rrnL+trnL+nad1*) resultante da inferência interespecífica com o programa *BEAST2*, a partir da metodologia apresentada no *ponto 3.5*. Esta foi calibrada com os parâmetros presentes na *Tabela 8 e* obtida a partir da combinação de 3 *runs* independentes.

Novamente continuando com a nomenclatura da inferência anterior vamos considerar os mesmos clades para a figura 13 (e figura 14) onde podemos começar com o intervalo observado no *clade A* onde temos um intervalo entre [16.1 , 16.367 Ma] com um valor *pp*=1 novamente suportando o artigo original em que o grupo Fulvus originou na Época Mioceno (A. Faille et al., 2014). Passando para o próximo clade observamos que temos um intervalo entre [7.5 , 9.07 Ma] com um suporte elevado (*pp*=1) no *clade B* dividindo assim os *clades C* e *D*, continuando assim para o *clade C* onde temos a separação entre a linhagem principal e a espécie iblanensis onde verificamos um intervalo entre [6.91 , 8.637 Ma] com um valor elevado de suporte (*pp*=1).

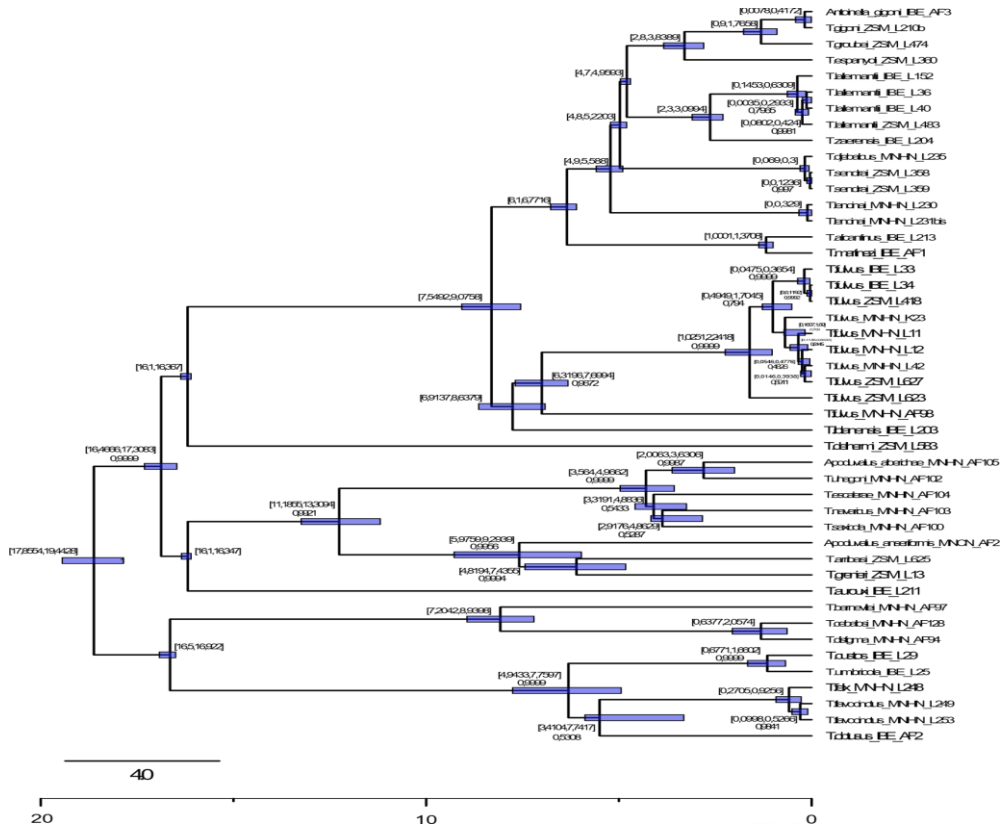


Figura 13 - Árvore filogenética resultante da inferência interespecífica ao dataset concatenado das escaravelhos do grupo Fulvus composto pelos genes COI, SSU, LSU e rrnL+trnL+nad1; calibrado com os parâmetros da Tabela 8 a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.

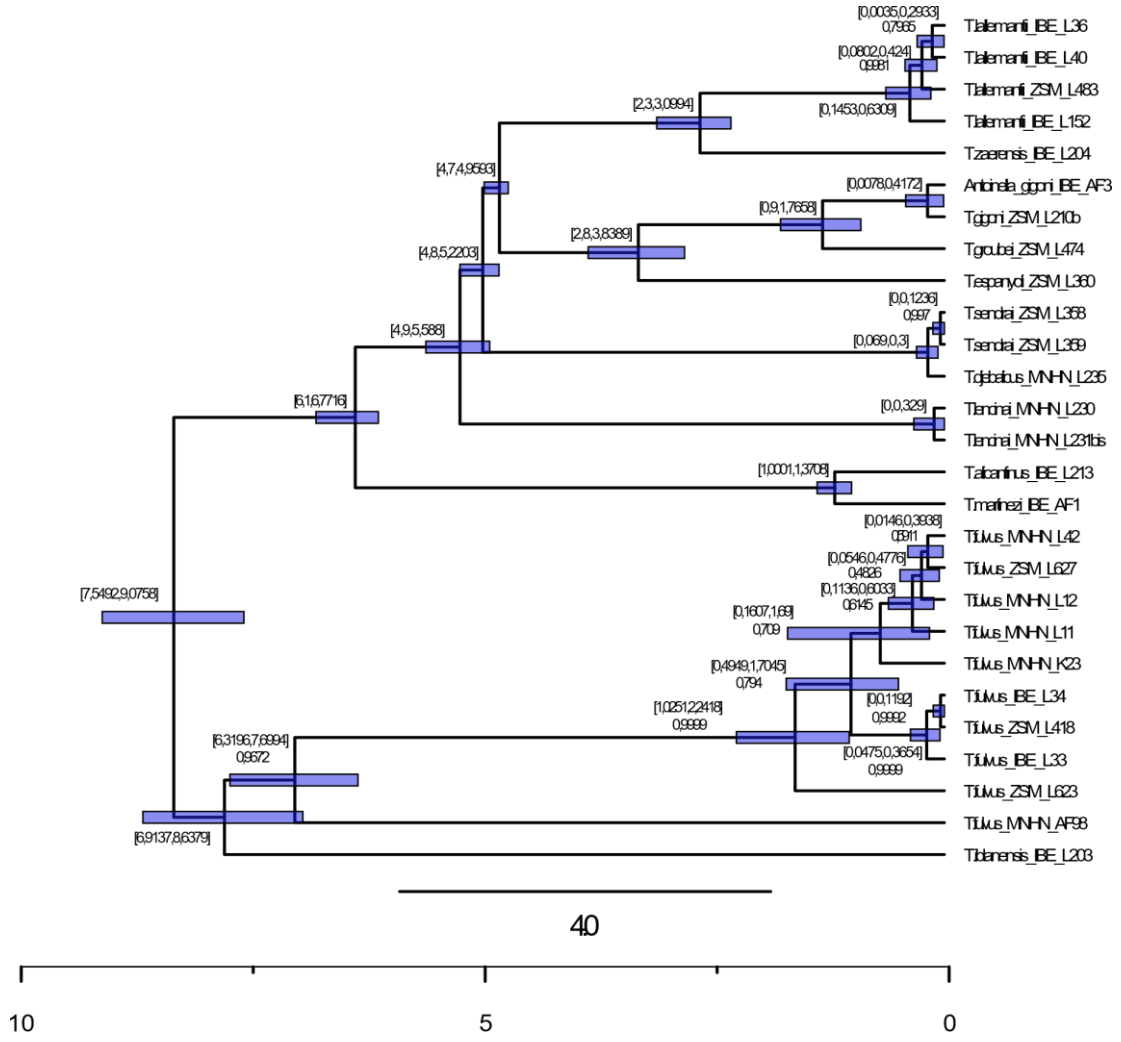


Figura 14 – Ampliação da Árvore filogenética resultante da inferência interespecífica ao dataset concatenado das escaravelhos do grupo Fulvus composto pelos COI, SSU, LSU e rrnL+trnL+nad1; calibrado com os parâmetros da Tabela 8 a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.

4.3.2 Inferência Intraespecífica, *BEAST 2

A *Figura 15 (e figura 16)* apresenta a árvore filogenética das espécies de escaravelhos do grupo Fulvus (*dataset* constituído pelos genes *COI, SSU, LSU e rrnL+trnL+nad1*) resultante da inferência interespecífica com o programa **BEAST2*, a partir da metodologia apresentada no *ponto 3.5*. Esta foi calibrada com os parâmetros presentes na *Tabela 9 e* obtida a partir da combinação de 3 *runs* independentes.

Com a nomenclatura da inferência anterior vamos considerar os mesmos clades para a figura 13 (e figura 14) onde podemos começar com o intervalo observado no *clade A* onde temos um intervalo entre [16.1 , 16.334 Ma] com um elevado valor *pp*=0.9544 suportando o artigo original em que o grupo Fulvus originou na Época Mioceno (A. Faille et al., 2014). Passando para o próximo clade observamos que temos um intervalo entre [7.1 , 8.06 Ma] com um suporte elevado (*pp*=1) no *clade B* dividindo assim os *clades C* e *D*, continuando assim para o *clade C* onde temos a separação entre a linhagem principal e a espécie iblanensis onde verificamos um intervalo entre [5.11 , 8.02 Ma] com um valor razoável de suporte (*pp*=0.9081). Em geral também se observa que ao longos dos clades os intervalos são cada vez mais curtos em *BEAST2 que em BEAST2 à medida que se aproxima do tempo presente.

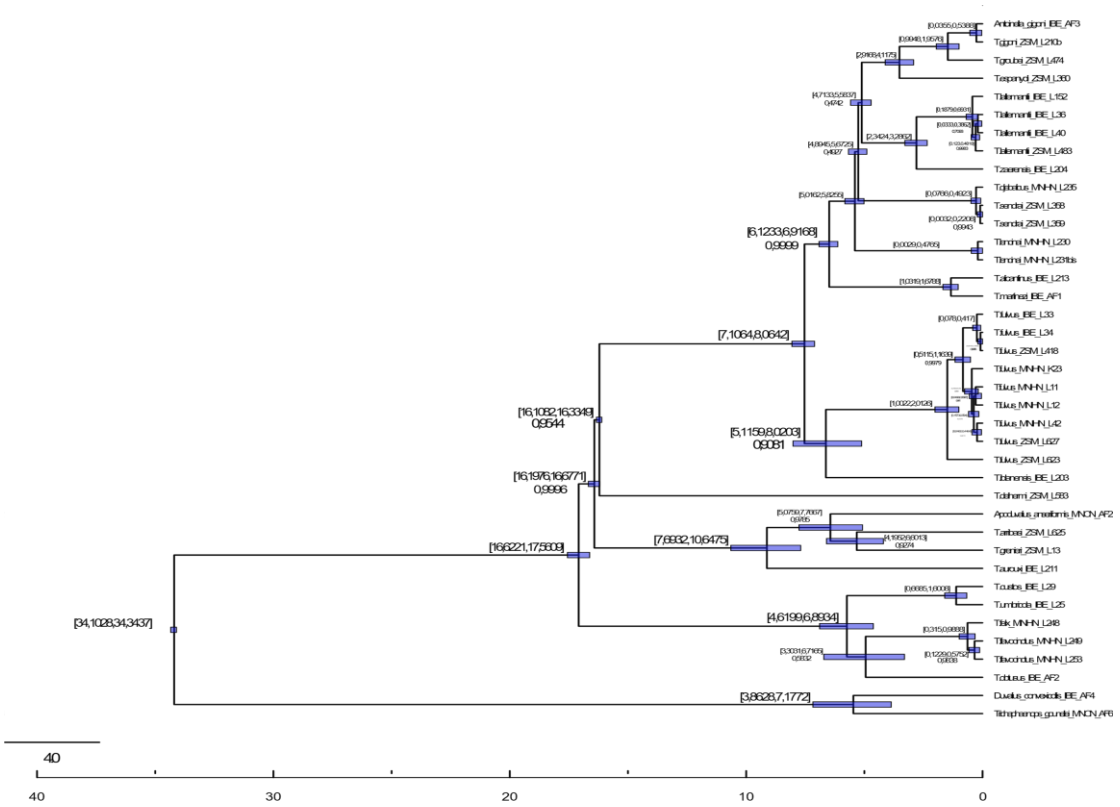


Figura 15 - Árvore filogenética resultante da inferência intraespecífica ao dataset concatenado dos Escaravelhos do grupo Fulvos composto pelos COI, SSU, LSU e rrnL+trnL+nad1; calibrado com os parâmetros da Tabela 9 e obtido a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.

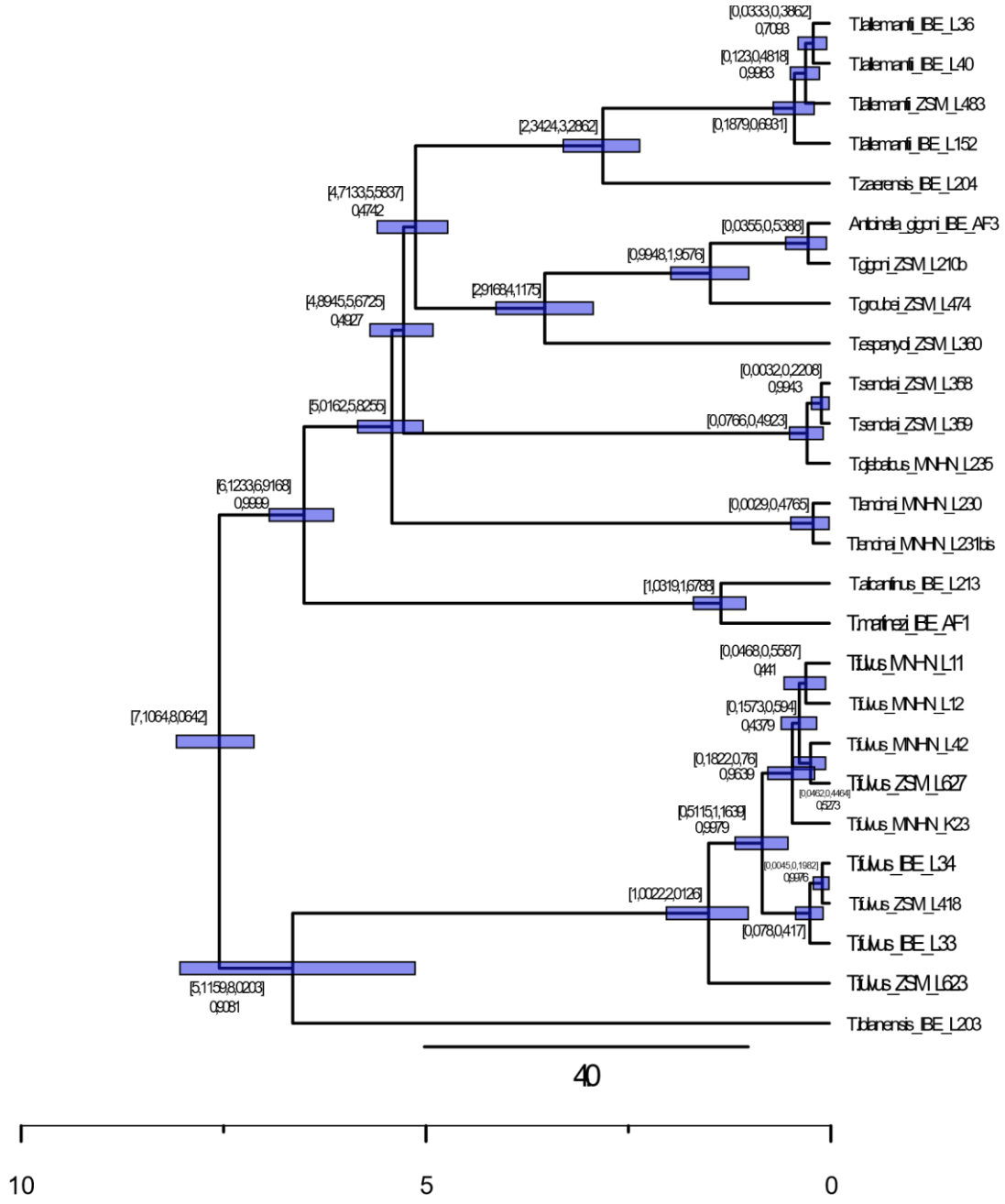


Figura 16 – Ampliação da Árvore filogenética resultante da inferência intraespecífica ao dataset concatenado dos Escaravelhos do grupo Fulvos composto pelos COI, SSU, LSU e rrnL+trnL+nad1; calibrado com os parâmetros da Tabela 9 e obtido a partir da combinação de 3 runs independentes. As barras azuis representam o HPD 95% com o valor em número próximo da barra e o valor pp sob o valor de HPD, se o valor de pp=1 não é representado no gráfico.

4.4 *Trechus fulvus* – BEAST2 (versão 2.4.0)

Na *Figura 17 (e figura 18)*, encontra-se apresentada a árvore filogenética de inferência interespecífica às espécies de escaravelhos do grupo Fulvus () com os genes COI, SSU, LSU e rrnL+trnL+nad1, obtida a partir uso da versão antiga do programa BEAST2 e *BEAST2 com combinação de 3 *runs* independentes e calibrada com os parâmetros da *Tabela* correspondente.

Considerando os mesmos clades para a figura 17 (e figura 18) onde podemos começar com o intervalo observado no *clade A* onde temos um intervalo entre [16.1 , 16.36 Ma] com um valor *pp*=1 colocando o grupo Fulvus com origem na Época Mioceno como se já havia visto nas inferências anteriores (A. Faille et al., 2014). Passando para o próximo clade observamos que temos um intervalo entre [7.55 , 9.08 Ma] com um suporte de valor *pp*=1 no *clade B* dividindo assim os *clades C* e *D*, seguimos assim para o *clade C* onde temos a separação entre a linhagem principal e a espécie iblanensis onde verificamos um intervalo entre [6.92 , 8.66 Ma] novamente com um valor *pp*=1.

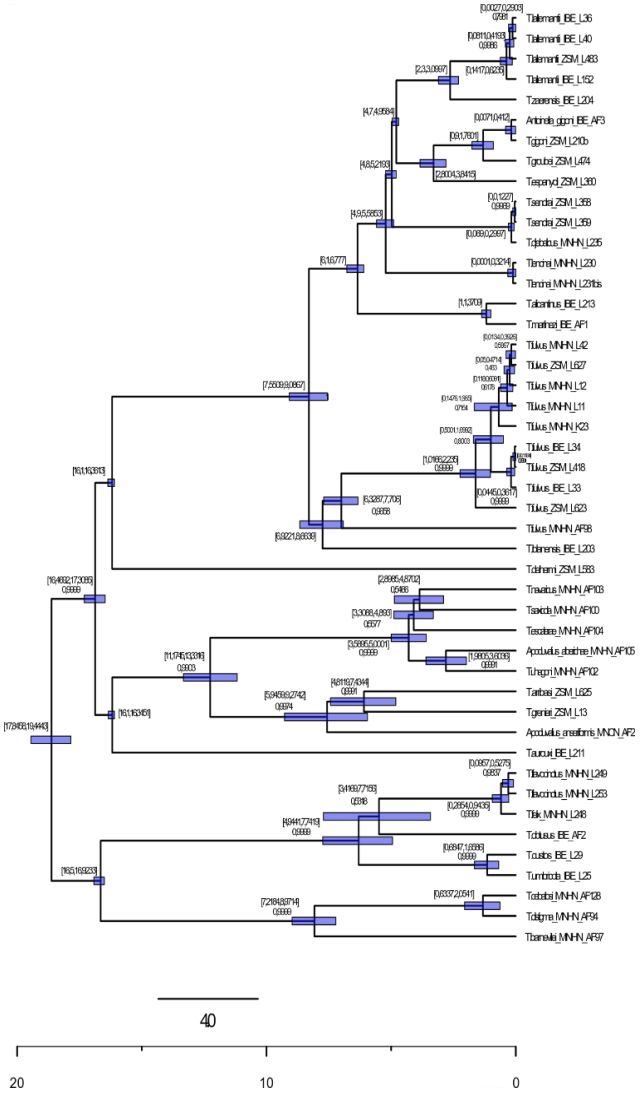


Figura 17 - Árvore filogenética resultante da inferência interespecífica ao dataset das escaravelhos do grupo Fulvus () composto pelos genes COI, SSU, LSU e rrnL+trnL+nad1; calibrado com os parâmetros da Tabela 8, obtido a partir da combinação de 3 runs independentes e do uso de uma versão antiga da aplicação BEAST2.

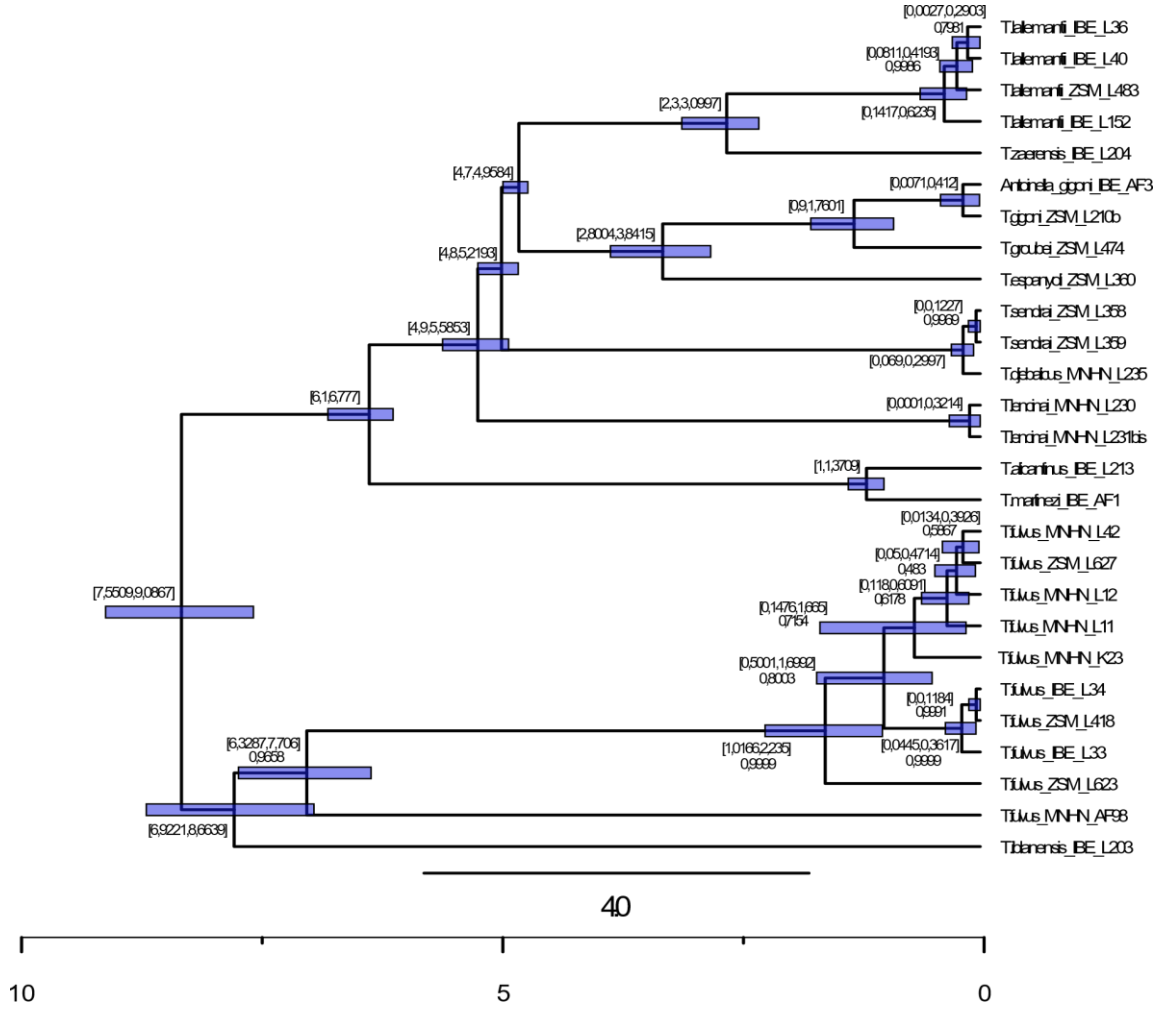


Figura 18 – Ampliação da Árvore filogenética resultante da inferência interespecífica ao dataset das escaravelhos do grupo Fulvus () composto pelos genes COI, SSU, LSU e rrnL+trnL+nad1; calibrado com os parâmetros da Tabela 8, obtido a partir da combinação de 3 runs independentes e do uso de uma versão antiga da aplicação BEAST2.

Na figura 19 com o método **BEAST2* verifica-se que a divergência no clade A poderá ter ocorrido no intervalo [16.1 , 16.38 Ma] com *pp*=0,9327, a separação no clade B no intervalo [7.10, 8.31 Ma], com um valor de suporte *pp*=0.9827 e a linhagem separa-se da espécie iblanensis no intervalo [4.98 , 8.52 Ma] com pp=0.8301.

Novamente com esta análise verifica-se as semelhanças entre os tempos de divergência e os valores de pp (posterior probability) independentemente da versão usada, levando à conclusão que a maior diferença entre versões se baseia no tempo de processamento da análise como se irá discutir na secção final.

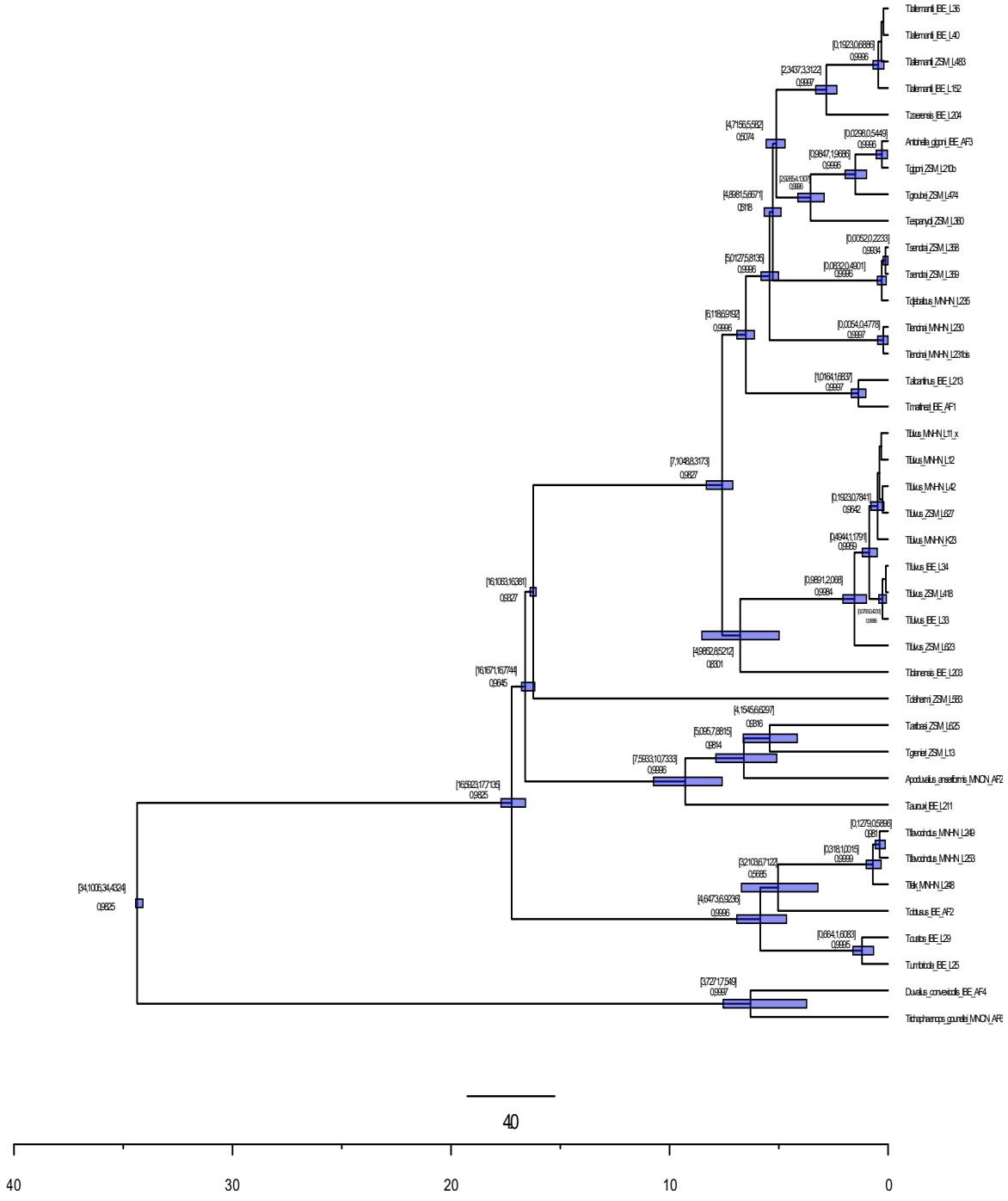


*Figura 19 – Árvore filogenética resultante da inferência intraspecífica ao dataset das escaravelhos do grupo Fulvus () composto pelos genes COI, SSU, LSU e rrnL+trnL+nad1; calibrado com os parâmetros da Tabela 9, obtido a partir da combinação de 3 runs independentes e do uso de uma versão antiga da aplicação *BEAST2.*

5 Discussão de resultados

Para esta discussão vão ser considerados como principais factores, os valores pp (Probabilidade posterior) representados nos nodes, os valores dos tempos de divergência assim como os efeitos do método de coalescência multi-espécies nas análises e o tempo de processamento de modo a podermos considerar qual dos métodos, *BEAST2* ou **BEAST2*, atualmente é o mais viável de ser aplicado tendo conta o suporte e a resolução da árvore filogenética. Em relação aos tipos de relógios moleculares foram usados os mesmos dados presentes nos artigos estudados para a sua reprodução assim como para obter resultados de maior fidelidade.

5.1 Espécies de escaravelhos tigre (Coleoptera: Cicindelini)

Apesar da análise dos valores de ESS no programa *Tracer* mostrar boa convergência dos parâmetros; os *Anexos 5* e *6* apresentam o resultado das distribuições dos parâmetros de ambas as abordagens. É importante realçar que o *dataset* é constituído por um amplo grupo de individuos/amostras suportado apenas pela inferência a partir de dois genes (*COI +16S*), o que faz com que a sua resolução possa não ser suficiente para inferir com confiança todas as relações filogenéticas entre os taxa presentes.

Dataset* de escaravelhos tigre – *COI +16S						
Principais Separações	Tempos de Divergência		*PP* no nodo de agrupamento		Tempo de Processamento (min)	
	BEAST2	**BEAST2*	*BEAST2*	**BEAST2*	*BEAST2*	**BEAST2*
Clade I	[2.2 , 2.65 Ma]	[3.66 , 4.58 Ma]	1	0.9907	58	33.3
Clade II	[1.4 , 1.95 Ma]	[2.14 , 3.2 Ma]	0.9999	0.9991		
Clade III	[0.65 , 1.15 Ma]	[0.89,2.3 Ma]	1	0.997		

Tabela 10 - Resultados dos tempos de divergência relativos, valores de pp no nodo de agrupamento e tempo de processamento das principais separações referentes à inferência filogenética molecular inter e intraespecífica, aos escaravelhos tigre (Coleoptera: Cicindelini) a partir do gene COI e do gene 16S. Especificações técnicas do computador onde foram executados os processos: Intel® Core™ n3540, NVIDIA 920m, 8 GB DDR3.

Com os resultados obtidos para o escaravelhos tigre observou-se que em termos de valor de pp ambos apresentam valores elevados (embora BEAST2 apresente valores maiores), para os tempos

de divergência com ambas as abordagens apresentados na *Tabela 10* demonstram que, apesar da diferença ser mínima, o *BEAST2* apresenta intervalos de tempo de divergência menores.

Por último, o método de inferência filogenética molecular **BEAST2* apresenta ser ~1,75 vezes mais rápido que o método *BEAST2* no que diz respeito ao tempo de processamento.

Estes resultados podem ser devidos ao facto do método MSC ser mais indicado para uso em datasets com várias espécies distintas o que não acontece com este exemplo em que embore houve várias amostras recolhidas, o número de espécies distintas não era elevado assim como apenas se usou dois genes para a análise, o que novamente não é o mais indicado para o método MSC em que uma variedade elevada de genes é preferível de modo a se poder lidar com eventuais ILCs.

5.2 *Trechus fulvus* - Dataset composto por COI, SSU, LSU e rrnL+trnL+nad1

Dataset de escaravelhos do grupo Fulvus – *COI, SSU, LSU , rrnL+trnL+nad1*						
Principais Separações	Tempos de Divergência		*PP* no nodo de agrupamento		Tempo de Duração de Processamento (min)	
	BEAST2	**BEAST2*	*BEAST2*	**BEAST2*	*BEAST2*	**BEAST2*
Clade A	[16.1 , 16.367 Ma]	[16.1 , 16.334 Ma]	1	0.9544	330	260
Clade B	[7.5 , 9.07 Ma]	[7.1 , 8.06 Ma]	1	1		
Clade C	[6.91 , 8.637 Ma]	[5.11 , 8.02 Ma]	1	0.9081		

Tabela 11 - Resultados dos tempos de divergência relativos, valores de pp no nodo de agrupamento e tempo de processamento das principais separações referentes à inferência filogenética molecular inter e intraespecífica, ao dataset dos escaravelhos do grupo Fulvus a partir do gene COI. Especificações técnicas do computador onde foram executados os processos: Intel® Core™ n3540, NVIDIA 920m, 8 GB DDR3.

De acordo com a *Tabela 11*, o método **BEAST2 e*m comparação com BEAST2 embora os valores dos tempos de divergência dos clades principais se possam considerar praticamente equivalentes, é de notar que mais perto do tempo presente os intervalos de tempo são em geral mais curtos que os vistos no método BEAST2, quanto aos valores de suporte pp ambos os métodos obtiveram resultados elevados embora BEAST2 tenha obtido os melhores valores de suporte deixando por fim o critério decisivo que é o tempo de processamento em que o *BEAST2 é aproximadamente 75% mais rápido que o BEAST2 o que também mostra a útilidade do método MSC tendo em conta que ao limitar as árvores de genes com a árvore de espécies torna mais restrito o "path" a seguir pelo algoritmo elevando a sua velocidade de processamento.

5.3 Diferenças entre as versões recentes(v2.6.3) e antigas(v2.4.0)

Após realizar as inferências filogenéticas do dataset de escaravelhos do grupo Fulvus em ambas as versões do programa BEAST2 e *BEAST2 averiguou-se que em termos de tempos de divergência e pp (posterior probability) não se verifica uma diferença significativa embora haja uma melhoria da versão antiga para a mais recente (especialmente nos valores de suporte pp no *BEAST2), a característica significativa entre as duas versões verifica-se na sua duração de processamento.
Ambas melhoraram de uma versão para a outra porém, com os specs deste PC (intelº core n3540, NVIDIA 920m e 8Gb DD3) as análises passaram de aproximadamente 3 horas para 1 hora e meia usando a BEAST2 e de 2 horas para aproximadamente meia hora com a *BEAST2.

6 Conclusões

Com todos os pontos possíveis considerados dentro das limitações dos modelos estudados, pode-se afirmar que existe uma melhoria com o uso de *BEAST2 sobre o uso de BEAST2, a diferença de intervalos de tempos de divergência e valores de suporte pp embora não tenha sido significativa na comparação entre análises para ser o factor de escolha entre uma plataforma ou outra, é de notar que o uso do método MSC foi dos factores que mais deixou impacto na velocidade de processamento na análise de *BEAST2 tendo em conta que o algoritmo tinha restrições no caminho a percorrer com o uso da árvore de espécies.

As vantagens claras para o uso de *BEAST2 ao invés de BEAST2 são a não obrigatoriedade de uso de um Outgroup, algo imprescindivel numa análise com BEAST2 para um bom suporte de confiança. A outra grande vantagem exercida pelo uso de *BEAST2 é o uso do método MSC (Multi species Coalescent Process) onde modelamos árvores de genes dentro das restrições da árvore de espécies resultando numa melhoria nos valores de suporte (posterior), assim como mitigando erros presentes em outros métodos como a presença de ILS (Incomplete Lineage Sorting).

A velocidade de processamento de análises com o programa *BEAST2 comparativamente às análises efetuadas pelo BEAST2 viu-se ser significativa em média apresentando uma velocidade 75% mais elevada o que em maiores datasets – tanto em número de taxa como de genes - irá fazer uma diferença significativa.

Contudo para certos tipos de análise, como foi visto nos escaravelhos tigre, onde foram usados apenas dois genes – COI e 16S – assim como apenas amostras recolhidas de vários individuos da mesma espécie, embora o método MSC tenha melhorado o tempo de processamento dessas análises, é de notar que tanto os intervalos de tempo de divergência assim como os valores de suporte pp, ainda que mínima, obtiveram resultados melhores (tendo em conta que valores iguais ou superiores a 0.95 são considerados como bons valores de suporte).

Concluíndo, a escolha entre o uso de BEAST2 e *BEAST vai depender mais do que utilizador pretende, qual o material a usar e qual o tempo disponível para as análises.

7 Referências Bibliográficas

[1] D. E. Soltis e P. S. Soltis, «The Role of Phylogenetics in Comparative Genetics»,
Plant Physiology, vol. 132, n. 4, pp. 1790–1800, Ago. 2003.

[2] M. R. Dietrich, «Paradox and Persuasion: Negotiating the Place of Molecular Evolution within Evolutionary Biology», *Journal of the History of Biology*, vol. 31, n. 1, pp. 85–111, Mar. 1998.

[3] P. Ajawatanawong, «Molecular Phylogenetics: Concepts for a Newcomer», *Adv. Biochem. Eng. Biotechnol.*, vol. 160, pp. 185–196, 2017.

[4] C. Senés-Guerrero, G. Torres-Cortés, S. Pfeiffer, M. Rojas, e A. Schüßler, «Potato-associated arbuscular mycorrhizal fungal communities in the Peruvian Andes», *Mycorrhiza*, vol. 24, n. 6, pp. 405–417, Ago. 2014.

[5] E. Kenah, T. Britton, M. E. Halloran, e I. M. L. Jr, «Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees», *PLOS Computational Biology*, vol. 12, n. 4, p. e1004869, Abr. 2016.

[6] A. B. Chang, R. Lin, W. Keith Studley, C. V. Tran, e M. H. Saier, «Phylogeny as a guide to structure and function of membrane transport proteins», *Mol. Membr. Biol.*, vol. 21, n. 3, pp. 171–181, Jun. 2004.

[7] J. B. H. Martiny, S. E. Jones, J. T. Lennon, e A. C. Martiny, «Microbiomes in light of traits: A phylogenetic perspective», *Science*, vol. 350, n. 6261, p. aac9323, Nov. 2015.

[8] M. Siljic *et al.*, «Forensic application of phylogenetic analyses - Exploration of suspected HIV-1 transmission case», *Forensic Sci Int Genet*, vol. 27, pp. 100–105, 2017.

[9] K. A. Jacobson, S. Costanzi, e S. Paoletta, «Computational studies to predict or explain GPCR polypharmacology», *Trends Pharmacol Sci*, vol. 35, n. 12, pp. 658– 663, Dez. 2014.

[10] S. Ojosnegros e N. Beerenwinkel, «Models of RNA virus evolution and their roles in vaccine design», *Immunome Res*, vol. 6, n. Suppl 2, p. S5, Nov. 2010.

[11] H. A. Khan, I. A. Arif, A. H. Bahkali, A. H. Al Farhan, e A. A. Al Homaidan, «Bayesian, maximum parsimony and UPGMA models for inferring the phylogenies of antelopes using mitochondrial markers», Evol. Bioinform. Online, vol. 4, pp. 263–270, Out. 2008.

[12] N. Saitou e M. Nei, «The neighbor-joining method: a new method for reconstructing phylogenetic trees», *Mol. Biol. Evol.*, vol. 4, n. 4, pp. 406–425, Jul. 1987.

[13] J. Wen, Y. Xu, Z. Li, Z. Ma, e Y. Xu, «Inter-class sparsity based discriminative least square regression», *Neural Netw*, vol. 102, pp. 36–47, Jun. 2018.

[14] S. Bastkowski, V. Moulton, A. Spillner, e T. Wu, «The minimum evolution problem is hard: a link between tree inference and graph clustering problems», *Bioinformatics*, vol. 32, n. 4, pp. 518–522, Fev. 2016.

[15] D. Ortega-Del Vecchyo, D. Piñero, L. Jardón-Barbolla, e J. van Heerwaarden,
«Appropriate homoplasy metrics in linked SSRs to predict an underestimation of demographic expansion times», *BMC Evolutionary Biology*, vol. 17, n. 1, p. 213, Set. 2017.

[16] C.-B. Stewart, «The powers and pitfalls of parsimony», *Nature*, vol. 361, n. 6413, pp. 603–607, Fev. 1993.

[17] E. L. Lawler e D. E. Wood, «Branch-and-Bound Methods: A Survey», *Operations Research*, vol. 14, n. 4, pp. 699–719, Ago. 1966.

[18] A. Goëffon, J.-M. Richer, e J.-K. Hao, «Heuristic Methods for Phylogenetic Reconstruction with Maximum Parsimony», em *Algorithms in Computational Molecular Biology*, John Wiley & Sons, Ltd, 2010, pp. 579–597.

[19] L. Kannan e W. C. Wheeler, «Maximum Parsimony on Phylogenetic networks»,
Algorithms Mol Biol, vol. 7, n. 1, p. 9, Mai. 2012.

[20] J. Felsenstein, «Evolutionary trees from DNA sequences: A maximum likelihood approach», *J Mol Evol*, vol. 17, n. 6, pp. 368–376, Nov. 1981.

[21] A. Som, «Theoretical foundation to estimate the relative efficiencies of the Jukes- Cantor+gamma model and the Jukes-Cantor model in obtaining the correct phylogenetic tree», *Gene*, vol. 385, pp. 103–110, Dez. 2006.

[22] M. Kimura, «A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences», *J Mol Evol*, vol. 16, n. 2, pp. 111–120, Jun. 1980.

[23] M. Hasegawa, H. Kishino, e T. Yano, «Dating of the human-ape splitting by a molecular clock of mitochondrial DNA», *J Mol Evol*, vol. 22, n. 2, pp. 160–174, Out. 1985.

[24] Degnan, James & Rosenberg, N.A.. (2009). Gene tree discordance, phylogenetic and the multispecies coalescent. Trends in Ecology & Evolution. 24. 332-340.

[25] K. Tamura e M. Nei, «Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees», *Mol. Biol. Evol.*, vol. 10, n. 3, pp. 512–526, Mai. 1993.

[26] L. Gatto, D. Catanzaro, e M. C. Milinkovitch, «Assessing the Applicability of the GTR Nucleotide Substitution Model Through Simulations», *Evol Bioinform Online*, vol. 2, pp. 145–155, Fev. 2007.

[27] P. Liò e N. Goldman, «Models of molecular evolution and phylogeny», *Genome Res.*, vol. 8, n. 12, pp. 1233–1244, Dez. 1998.

[28] J. Sullivan e P. Joyce, «Model Selection in Phylogenetics», *Annual Review of Ecology, Evolution, and Systematics*, vol. 36, n. 1, pp. 445–466, 2005.

[29] M. Arenas, «Trends in substitution models of molecular evolution», *Front Genet*, vol. 6, Out. 2015.

[30] D. Posada e K. A. Crandall, «MODELTEST: testing the model of DNA substitution.», *Bioinformatics*, vol. 14, n. 9, pp. 817–818, Jan. 1998.

[31] D. Darriba, G. L. Taboada, R. Doallo, e D. Posada, «jModelTest 2: more models, new heuristics and parallel computing», *Nature Methods*, vol. 9, n. 8, pp. 772–772, Ago. 2012.

[32] J. Bertl, G. Ewing, C. Kosiol, e A. Futschik, «Approximate maximum likelihood estimation for population genetic inference», *Stat Appl Genet Mol Biol*, vol. 16, n. 5–6, pp. 387–405, 27 2017.

[33] Z. Yang e B. Rannala, «Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds», *Mol. Biol. Evol.*, vol. 23, n. 1, pp. 212–226, Jan. 2006.

[34] M. E. Alfaro e M. T. Holder, «The Posterior and the Prior in Bayesian Phylogenetics», *Annual Review of Ecology, Evolution, and Systematics*, vol. 37, n. 1, pp. 19–42, 2006.

[35] M. Holder e P. O. Lewis, «Phylogeny estimation: traditional and Bayesian approaches», *Nat. Rev. Genet.*, vol. 4, n. 4, pp. 275–284, Abr. 2003.

[36] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, e A. Stamatakis, «RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference», *Bioinformatics*, vol. 35, n. 21, pp. 4453–4455, Nov. 2019.

[37] J. P. Huelsenbeck e F. Ronquist, «MRBAYES: Bayesian inference of phylogenetic trees», *Bioinformatics*, vol. 17, n. 8, pp. 754–755, Ago. 2001.

[38] F. Ronquist *et al.*, «MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space», *Syst. Biol.*, vol. 61, n. 3, pp. 539–542, Mai. 2012.

[39] R. Bouckaert *et al.*, «BEAST 2: A Software Platform for Bayesian Evolutionary Analysis», *PLOS Computational Biology*, vol. 10, n. 4, p. e1003537, Abr. 2014.

[40] G. J. Morgan, Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965», *Journal of the History of Biology*, vol. 31, n. 2, pp. 155–178, 1998.

[41] L. Bromham e D. Penny, «The modern molecular clock», *Nature Reviews Genetics*, vol. 4, n. 3, pp. 216–224, Mar. 2003.

[42] M. S. Y. Lee e S. Y. W. Ho, «Molecular clocks», *Current Biology*, vol. 26, n. 10, pp. R399–R402, Mai. 2016.

[43] M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, e A. Rambaut,
«Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10»,
Virus Evol, vol. 4, n. 1, p. vey016, Jan. 2018.

[44] J. Barido-Sottani *et al.*, «Taming the BEAST-A Community Teaching Material Resource for BEAST 2», *Syst. Biol.*, vol. 67, n. 1, pp. 170–174, 01 2018.

[45] J. Heled e A. J. Drummond, «Bayesian inference of species trees from multilocus data», *Mol. Biol. Evol.*, vol. 27, n. 3, pp. 570–580, Mar. 2010.

[46] T. A. Heath, D. S. Divergences, e F. B. Process, «Divergence Time Estimation using BEAST v2.2.0 Dating Species Divergences with the Fossilized Birth-Death Process», n. Mcmc, pp. 1–44, 2016.

[47] Z. Yang e B. Rannala, «Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds», *Mol. Biol. Evol.*, vol. 23, n. 1, pp. 212–226, Jan. 2006.

[48] R. Bouckaert *et al.*, «BEAST 2.5: An advanced programa platform for Bayesian evolutionary analysis», *PLOS Computational Biology*, vol. 15, n. 4, p. e1006650, Abr. 2019.

[49] A. Rambaut, A. J. Drummond, D. Xie, G. Baele, e M. A. Suchard, «Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7», *Syst. Biol.*, vol. 67, n. 5, pp. 901–904, 01 2018

[50] A. Rambaut, «FigTree, version 1.4.3», 2009.

[51] K. Katoh e D. M. Standley, «MAFFT Multiple Sequence Alignment Programa Version 7: Improvements in Performance and Usability», *Mol Biol Evol*, vol. 30, n. 4, pp. 772–780, Abr. 2013.

[52] A. Larsson, «AliView: a fast and lightweight alignment viewer and editor for large datasets», *Bioinformatics*, vol. 30, n. 22, pp. 3276–3278, Nov. 2014.

[53] Santos,X. et al. «Complex phylogeography in the Southern Smooth Snake (Coronella girondica) supported by mtDNA sequences», Journal of Zoological Systematics and Evolutionary Research, 50, 210–219, 2012.

[54] I. Ribera *et al.*, «Ancient origin of a Western Mediterranean radiation of subterranean beetles», *BMC Evolutionary Biology*, vol. 10, n. 1, p. 29, Jan. 2010.

[55] E. Sérusiaux, J. C. V. A, T. Wheeler, e B. Goffinet, «Recent origin, active speciation and dispersal for the lichen genus Nephroma (Peltigerales) in Macaronesia», *Journal of Biogeography*, vol. 38, n. 6, pp. 1138–1151, 2011.

[56] I. R. Amorim, B. C. Emerson, P. A. V. Borges, e R. K. Wayne, «Phylogeography and molecular phylogeny of Macaronesian island Tarphius (Coleoptera: Zopheridae): why are there so few species in the Azores?», *Journal of Biogeography*, vol. 39, n. 9, pp. 1583–1595, 2012.

[57] A. Faille, C. Andújar, F. Fadrique, e I. Ribera, «Late Miocene origin of an Ibero- Maghrebian clade of ground beetles with multiple colonizations of the subterranean environment», *Journal of Biogeography*, vol. 41, n. 10, pp. 1979– 1990, 2014.

[58] T. Menezes, M. M. Romeiras, M. M. de Sequeira, e M. Moura, «Phylogenetic relationships and phylogeography of relevant lineages within the complex Campanulaceae family in Macaronesia», *Ecology and Evolution*, vol. 8, n. 1, pp. 88–108, 2018.

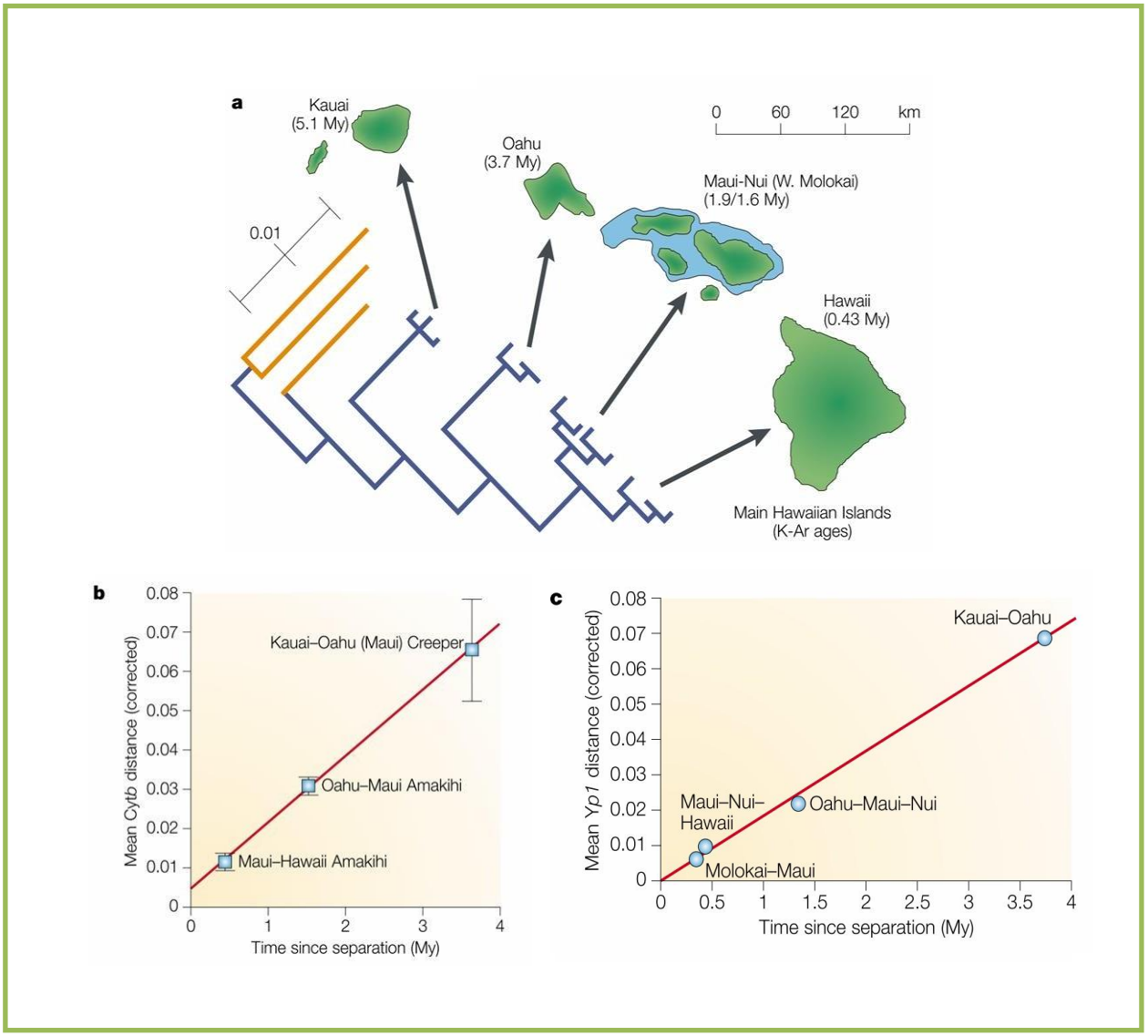
[59] García-Reina,A. et al., «Phylogeographic patterns of two tiger beetle species at both sides of the strait of Gibraltar (Coleoptera: Cicindelini)», Annales de la Société entomologique de France (N.S.), 50, 399–406, 2014.

[60] M. Espeland *et al.*, «A Comprehensive and Dated Phylogenomic Analysis of Butterflies», *Current Biology*, vol. 28, n. 5, pp. 770-778.e5, Mar. 2018.

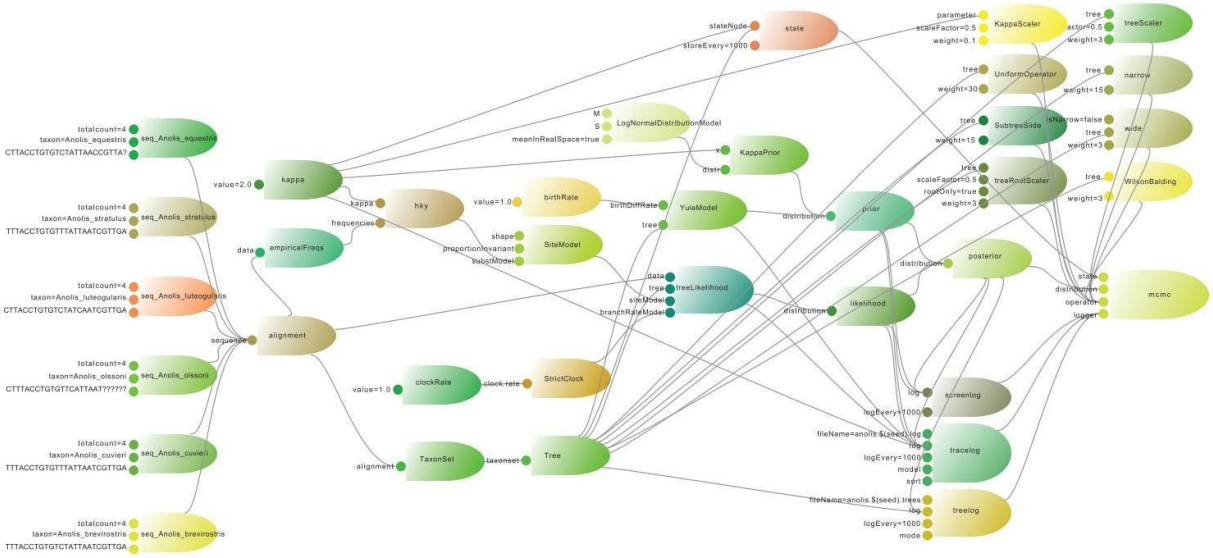
[61] C. Peña, S. Nylin, e N. Wahlberg, «The radiation of Satyrini butterflies (Nymphalidae: Satyrinae): a challenge for phylogenetic methods», *Zoological Journal of the Linnean Society*, vol. 161, n. 1, pp. 64–87, 2011.

[62] Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European journal of phycology, 49(2), 179-196.

8 Anexos



*Anexo 1 - Relógios moleculares das ilhas Havaianas. a**:** A origem vulcânica das ilhas Havaianas ocorreu com a formação de uma cadeia de ilhas com idades geológicas crescestes. As relações filogenéticas de aves insulares endémicas (como por exemplo espécies de drepananina: amakihi, Hemignathus virens e akiapolaau Hemignathus wilsoni; e moscas da fruta (Drosophila spp.)) refletem a "conveyer belt", com as espécies das primeiras ilhas mais antigas formarem os ramos mais internos da árvore evolutiva, e as ilhas mais recentes nas pontas. As "branches" as laranjas representam os outgroups. b,c : Datas moleculares para a espécie Hemignathus (**b**) e Drosophila (**c**) confirmam a ordem de colonização, e infere uma relação linear entre as distâncias da divergência genética e o tempo de separação em relação à idade da ilha. My, "million years" (milhões de anos); Fonte (L. Bromham e D. Penny, 2003) reproduzidas com a permissão da REF. 10 © (1998) Blackwell Publishing.*



Anexo 2 - Workflow com seis sequências, tendo por uso o modelo de substituição HKY, strict clock, Yule tree prior, e ficheiros de logs que produzem o output para o Tracer, a árvore e o "screen output" através do processo da MCMC. doi:10.1371/journal.pcbi.1003537.g001. Fonte (J. Heled e A. J. Drummond, 2010)

Materiais de estudo

http://beast.community/tree_priors

http://beast.community/rates_and_dates

http://beast.community/tip_dates

http://beast.community/treeannotator

http://beast.community/analysing_beast_output

https://github.com/Taming-the-BEAST/Substitution-model-averaging

https://journals.plos.org/ploscompbiol/article/file?id=info%3Adoi/10.1371/journal.pcbi.1003537.s004&type=supplementary

https://beast.community/second_tutorial

https://beast.community/tempest_tutorial

https://beast.community/phylodynamics_of_epidemic_influenza

https://beast.community/continuous_traits

https://beast.community/taxon_sets

https://beast.community/tip_dates

https://beast.community/tip_date_sampling

https://beast.community/constructing_models

https://beast.community/analysing_beast_output

https://beast.community/tracer_convergence

https://beast.community/markov_jumps_rewards

https://beast.community/adaptive_mcmc

https://beast.community/model_averaging_clocks

https://beast.community/time_dependent_rate_model

https://taming-the-beast.org/tutorials/StarBeast-Tutorial/

Anexo 3 – Lista de links dos materiais de estudo principais usados.

Título da Publicação	Ref.	Ano de Publicação	Espécies de Estudo	Localização	Escala de Tempo
"Complex phylogeography in the Southern Smooth Snake (Coronella Girondica) supported by mtDNA sequences"	(Santos,X. et al., 2012)	2012	*Cobra Bordalesa*	Ocidente do Mediterrâneo	Milhões de anos (Ma)
"Ancient origin of a Western Mediterranean radiation of subterranean beetles"	(I. Ribera et al., 2010)	2010	Escaravelhos subterraneos da tribo *Leptodirini (Coleoptera, Leiodidae, Cholevinae)*	Cordilheiras do Ocidente Mediterrâneo	Milhões de anos (Ma)
"Recent origin, active speciation and dispersal for the lichen genus Nephroma (Peltigerales) in Macaronesia"	(E. Sérusiaux et al., 2011)	2011	*Gene Nephroma (Peltigerales) do Lichen*	Cosmopolitano com foco nas ilhas da Macaronésia: Açores, Madeira, Ilhas Canárias	Milhões de anos (Ma)
"Phylogeography and molecular phylogeny of Macaronesian island Tarphius (Coleoptera: Zopheridae): why are there so few species in the Azores?"	(I. R. Amorim et al., 2012)	2012	*Tarphius (Coleoptera:Zopheridae)*	Macaronésia: Açores, Madeira, Ilhas Canárias	Milhões de anos (Ma)
"Late Miocene origin of an Ibero-Maghrebian clade of ground beetles with multiple colonizations of the subterranean environment"	(A. Faille et al., 2014)	2014	*Escaravelhos da espécie*	Paleártico Ocidental, com foco na área entre o sudeste da Península Ibérica e o Norte de Marrocos.	Milhões de anos (Ma)
"Phylogenetic relationships and phylogeography of relevant lineages within the complex Campanulaceae family in Macaronesia"	(T. Menezes et al., 2018)	2017	Família das *Campanulaceae*	Macaronésia: Açores, Madeira, Ilhas Canárias	Milhões de anos atrás (Ma)

Anexo 4 – Lista de publicações estudadas com adequação ao projeto.

Statistic	Mean	ESS	Type
posterior	-5362.083	15658	R
likelihood	-5310.632	5874	R
prior	-51.451	10854	R
treeLikelihood.16S	-1947.252	4038	R
treeLikelihood.coi	-3363.379	15762	R
TreeHeight.t:coi+16S	1.282	53592	R
clockRate	4.289E-2	4404	R
popSize.t:coi+16S	1.544	56863	R
CoalescentConstant.t:coi+16S	-51.033	10895	R
mrca.age(Cephalota Outgroup)	1.023	40896	R

Anexo 5 - Resultado dos valores de convergência dos parâmetros usados na calibração da MCMC do dataset dos escaravelhos tigre, analisados pelo programa *Tracer; executado com BEAST2.*

Statistic	Mean	ESS	Type
posterior	-5352.277	4016	R
likelihood	-5308.692	3472	R
prior	-2.642	3616	R
speciescoalescent	-40.943	1827	R
popMean	0.749	2063	R
TreeHeight.Species	0.992	2625	R
TreeHeight.t:coi+16S	1.298	9643	R
treeLikelihood.16S	-1947.423	2903	R
treeLikelihood.coi	-3361.269	6081	R
treePrior.t:coi+16S	-25.014	2925	R
clockRate.c:16S	4.021E-2	3427	R
popSize.t:Species	1.357	6226	R
CoalescentConstant.t:Species	-3.013	3237	R

Anexo 6 – Resultado dos valores de convergência dos parâmetros usados na calibração da MCMC do dataset dos escaravelhos tigre, analisados pelo programa *Tracer; executado com *BEAST2.*

9 Apêndices

```
# script efetch

wget https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi/\?db\=nucleotide\&id\= "ACCESSION NUMBERS DO NCBI SEPARADOS POR VIGULAS"\&rettype\=fasta -O ~/"PATH PARA A DIRETORIA DO SAVE".fasta
```

Apêndice 1 - Script efetch que faz o download dos alinhamentos das samples em formato fasta.