UNIVERSIDADE DE SÃO PAULO - IB - DEPARTAMENTO DE BOTÂNICA
BIB-5772 - BIOGEOGRAFIA DE PLANTAS VASCULARES - 2018
Prof. Dr. José Rubens Pirani - responsável

Estudo dirigido: Estimação de tempos de divergência[1]
Responsável: Profa. Dra. Maria Fernanda Calió (mfecalio@unicamp.br)

Introdução

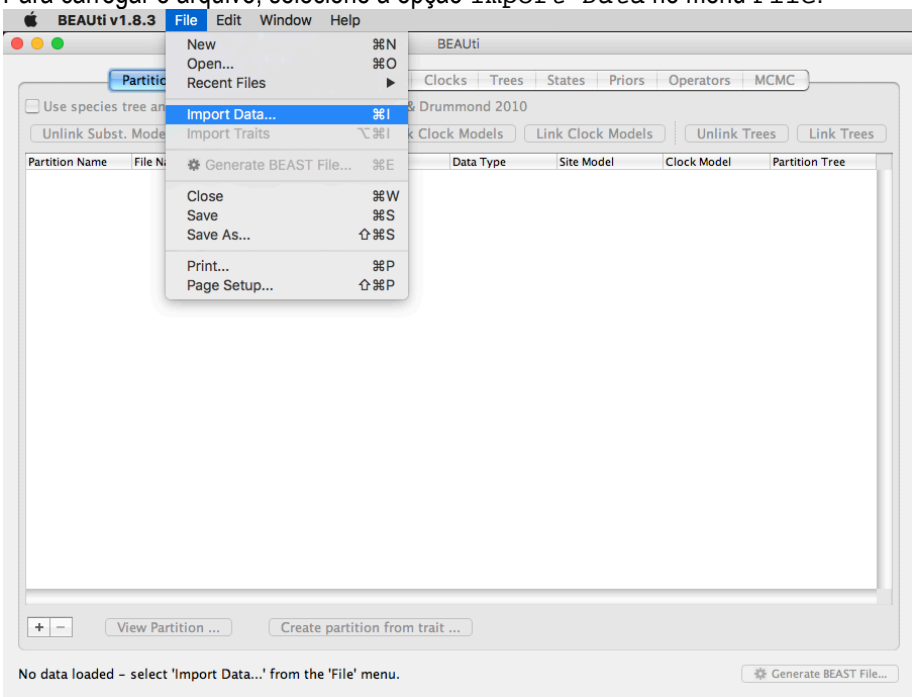
Essa atividade prática introduzirá o programa BEAST (*Bayesian Evolutionary Analysis Sampling Trees*) para análises evolutivas Bayesianas, com foco na estimação de filogenias e tempos de divergência quando você tem informações para calibração provenientes de evidência fóssil ou de outro conhecimento prévio (*prior*). Você precisará dos seguintes programas (disponíveis para *download* a partir de http://beast.community/index.html[2]):

- BEAST - esse pacote contém o programa BEAST, BEAUti, TreeAnnotator e outros utilitários.
- Tracer - esse programa é empregado para explorar os resultados da análise com o BEAST (e outros programas baseados em MCMC Bayesianos). Ele sumariza graficamente e quantitativamente as distribuições de parâmetros contínuos e provê informação diagnóstica.
- FigTree - esse é um aplicativo para visualização de filogenias moleculares, em particular aquelas obtidas usando BEAST.

Esse tutorial guiará você através da análise de um alinhamento de sequências de DNA de 12 espécies de primata. O objetivo é estimar a filogenia, bem como a taxa de evolução em cada linhagem baseada em datas de divergência. O primeiro passo é converter um arquivo NEXUS com os característicos blocos de dados ou de caracteres em um arquivo BEAST XML, usando o programa BEAUti (*Bayesian Evolutionary Analysis Utility*) para definir o modelo evolutivo e opções para a análise de MCMC. O segundo passo é realmente executar o BEAST usando o arquivo produzido com o BEAUTi (*input*), contendo os dados, modelos e configurações. O terceiro passo é explorar o resultado (*output*) do BEAST para diagnosticar potencias problemas. E os últimos passos são sumarização de dados e posterior visualização da árvore.

PASSO 1: BEAUTi

(1) Execute o BEAUti, clicando em seu ícone.
(2) Carregue o arquivo NEXUS.
• Para carregar o arquivo, selecione a opção `Import Data` no menu `File`:



[1] Traduzido e adaptado de **"Relaxed Phylogenetics and Dating with Confidence"**, de Drummond, Rambaut and Xie, 2012. O documento original é fornecido juntamente com o programa BEAST.
[2] Excelente documentação do programa e seus utilitários; vale à pena uma investigação minuciosa de seu conteúdo.

- Selecione o arquivo chamado primates.nex. Esse arquivo contém um alinhamento de sequências de 12 espécies de primatas e tem a seguinte aparência (quando examinado em um editor de texto):

primates.nex

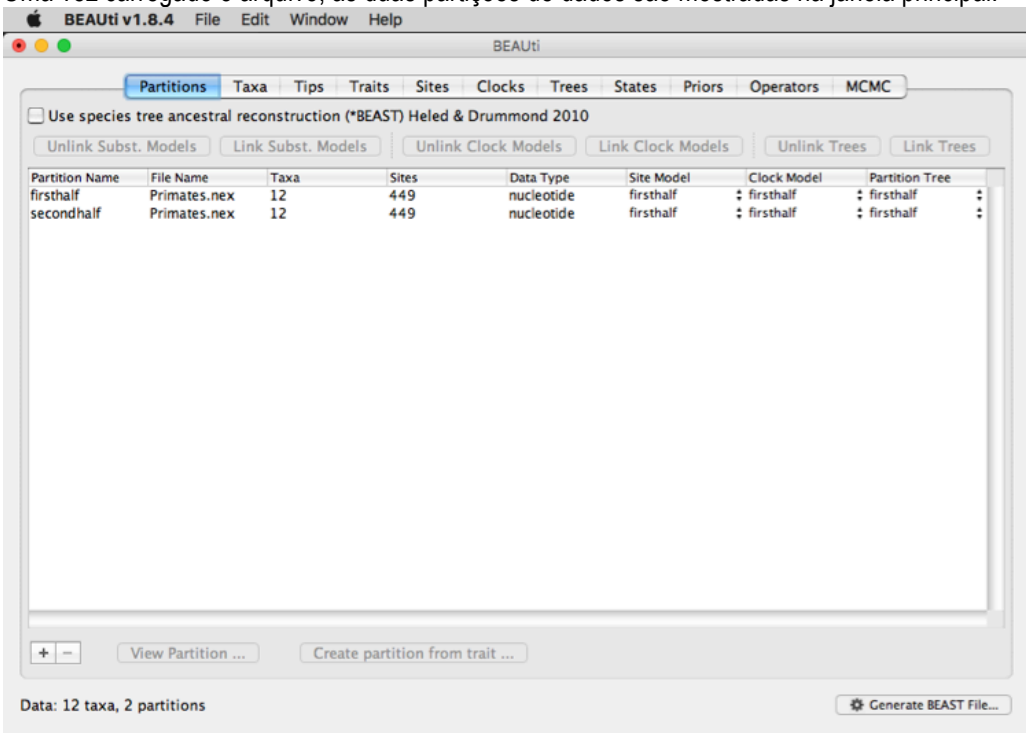
File Path ▾ : ~/Desktop/BEAST tutorial/primates.nex

primates.nex

```
#NEXUS
begin data;
dimensions ntax=12 nchar=898;
format datatype=dna interleave=no gap=-;
matrix
Tarsius_syrichta    AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCCTATTATTTTGCCTAGC
Lemur_catta     AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCATATTATTCTGTCTAGCCAAC
Homo_sapiens        AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTACTATTCTGCCTAGC
Pan         AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATTCTGCCTAGCAAACTCAA
Gorilla         AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATTCTGCCTAGCAAAC
Pongo           AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTTCTGCCTAGCAAAC
Hylobates       AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCTGCTATTCTGCCTTGCAAAC
Macaca_fuscata      AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCATATATTTCTGCCTAGC
M_mulatta       AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCATATATTTCTGCCTAGCCAAT
M_fascicularis      AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCATGTATTTCTGCTTGGC
M_sylvanus      AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCATATACTTCTGCTTGGCCAAC
Saimiri_sciureus    AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTATGCTATTCTGCCTAGC
;
end;

begin assumptions;
    charset firsthalf = 1-449;
    charset secondhalf = 450-898;
end;
```

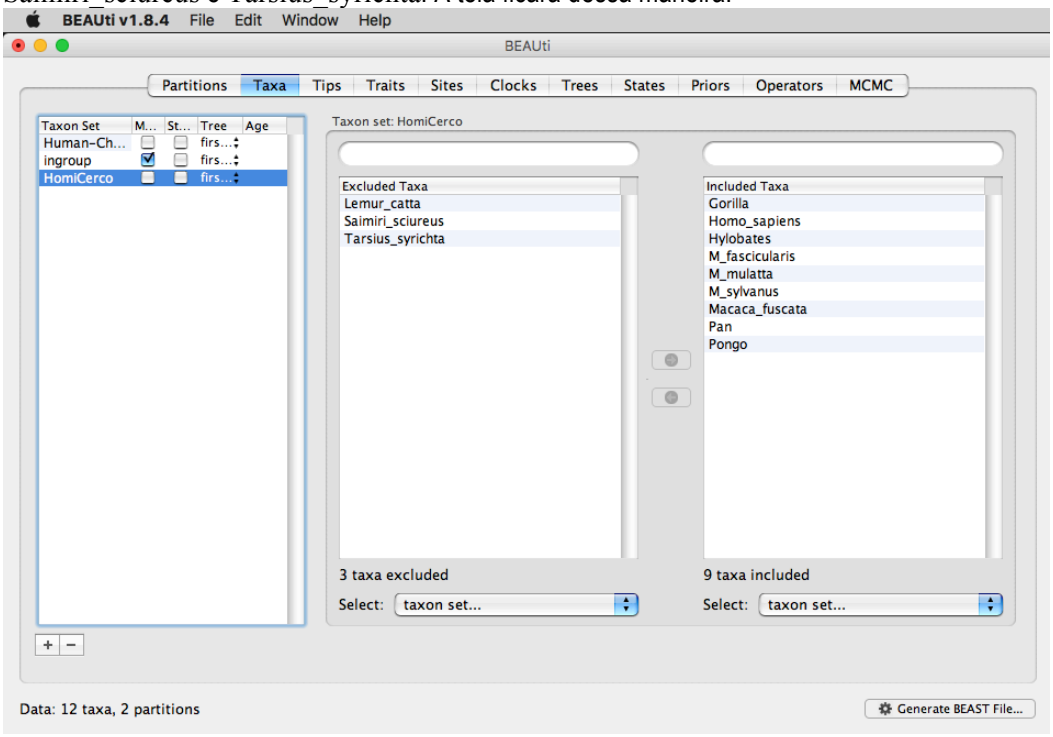
- Uma vez carregado o arquivo, as duas partições de dados são mostradas na janela principal:



(3) Defina os pontos de calibração.

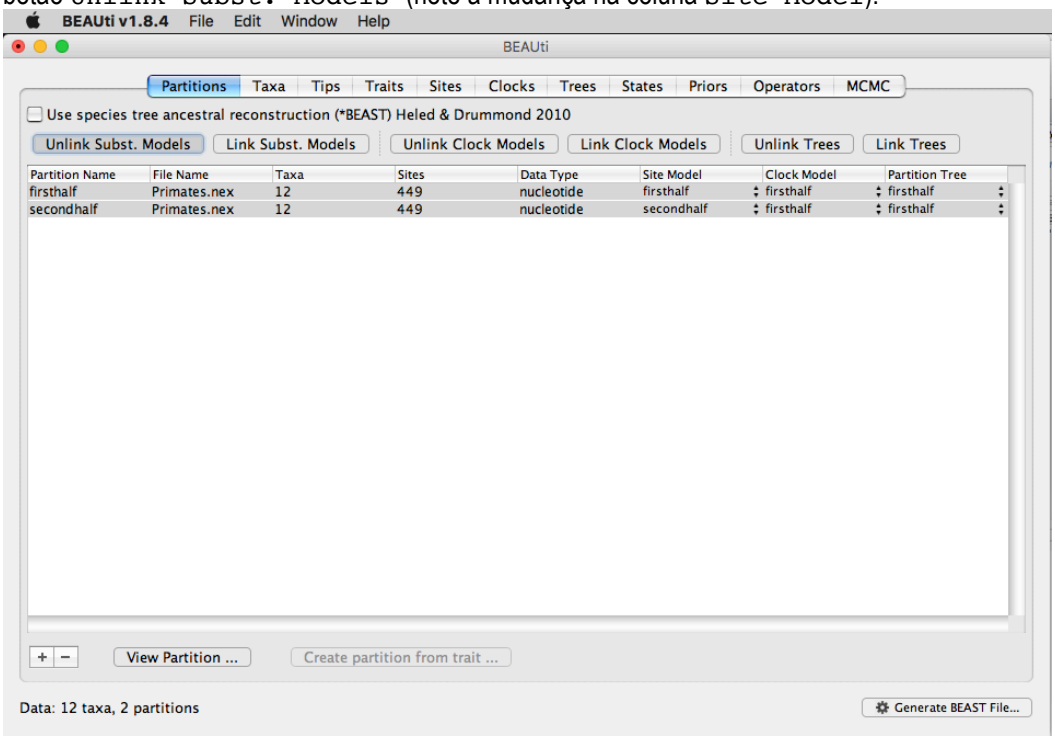
- Selecione a aba Taxa no topo da janela principal. Você verá o painel que permite criar conjuntos de táxons. Uma vez que você tiver criado um conjunto de táxons, você poderá adicionar informações de calibração para o seu ancestral comum mais recente (MRCA, *most recent common ancestor*).
- Pressione o botão "+" na porção inferior esquerda da janela. Isso criará um novo conjunto de táxons. Com um duplo-clique sobre o nome que aparece (*untitled1*), renomeie o conjunto, digitando "*ingroup*".
- À direita dessa coluna, você verá todos os táxons disponíveis. Selecione todos os táxons e pressione o botão com a seta verde. Mova Lemur_catta de volta para os táxons excluídos. Uma vez que sabemos que Lemur_catta é o grupo-externo, na coluna à esquerda, selecione o *checkbox* da coluna "Mono?". Isso assegurará que o grupo-interno seja mantido como monofilético durante o curso da análise de MCMC.
- Agora, crie um conjunto chamado "*Human-Chimp*" que contenha apenas os táxons Homo_sapiens e Pan.

- Finalmente, crie um grupo de táxons chamado "*HomiCerco*" que contenha todos os táxons, exceto Lemur_catta, Saimiri_sciureus e Tarsius_syrichta. A tela ficará dessa maneira:



(4) Desvincule modelos das partição (*unlink partitions models*).

- Neste ponto, será necessário desvincular o modelo de substituição para que cada parâmetro seja estimado separadamente para as duas partições. Para fazer isto, retorne à aba `Partitions`, selecione ambas as partições na tabela e clique no botão `Unlink Subst. Models` (note a mudança na coluna `Site Model`):

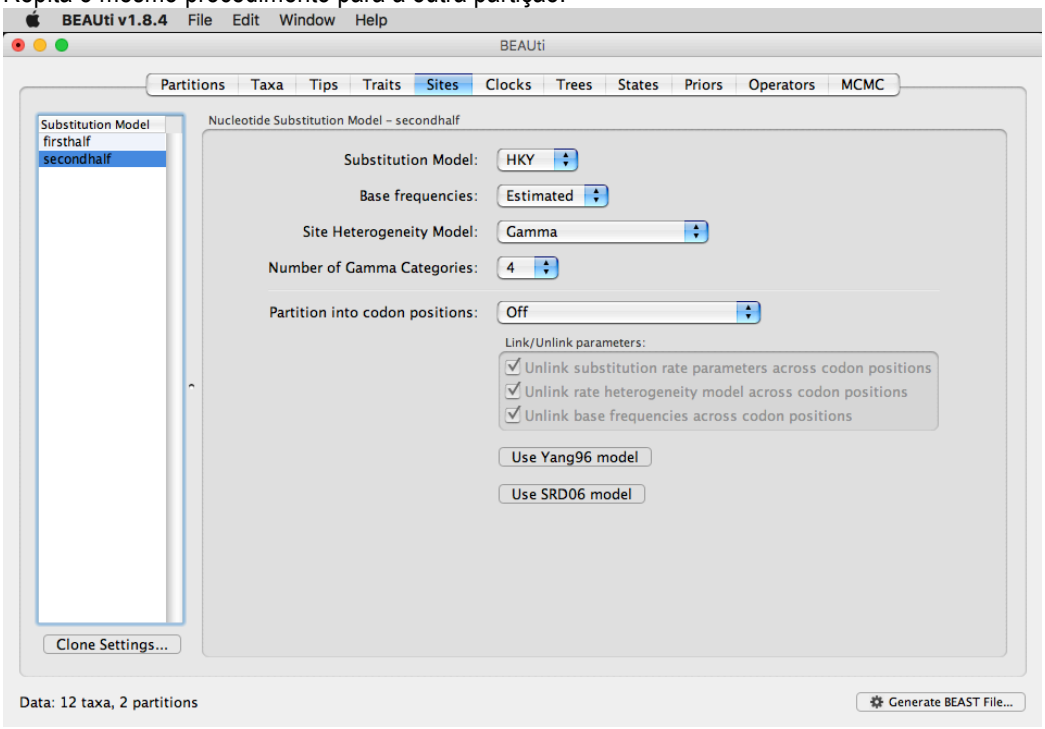


(5) Defina o modelo de substituição (*substitution model*).

- A próxima coisa a fazer é clicar na aba `Sites` no topo da janela principal. Nessa aba, você fará as configurações dos modelos evolutivos para o BEAST. As opções disponibilizadas dependerão do tipo de dados em uso (se são nucleotídeos,

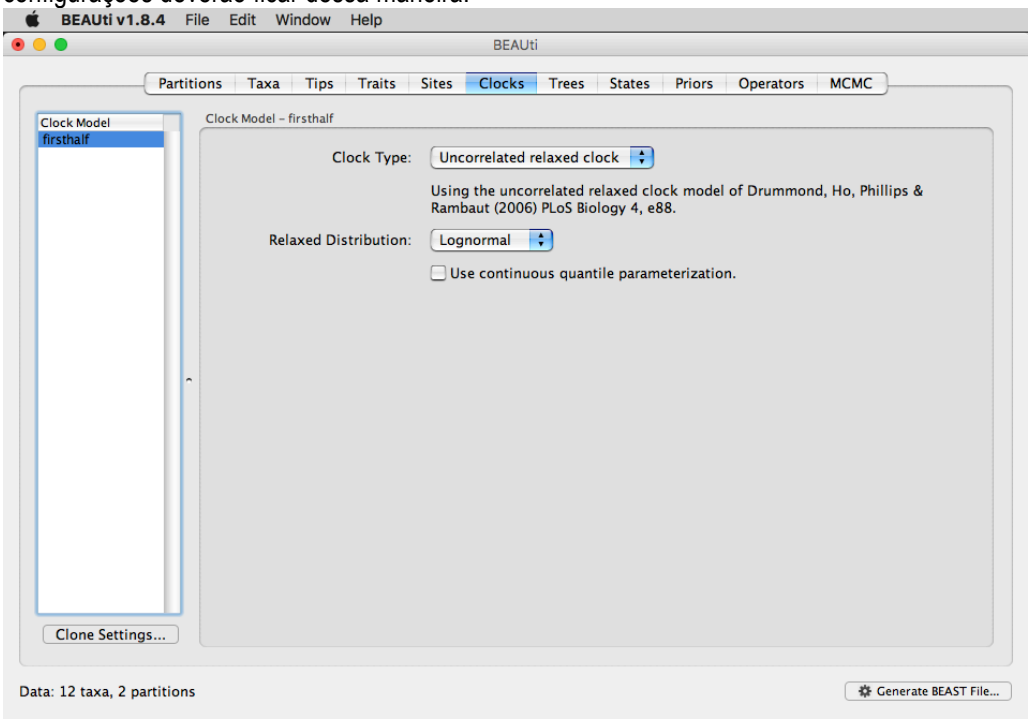
aminoácidos, dados binários). Após carregar qualquer arquivo para o BEAUti, as configurações *default* (padrão) irão aparecer, por isso faça uma avaliação cuidadosa, e se necessário, faça mudanças. Não se esqueça de que as alterações devem ser feitas separadamente para cada partição de dados.

- Para esta análise, na coluna da esquerda `Susbtitution Model`, selecione uma das partições, e na porção à direita, em `Site Heterogeneity Model`, selecione "Gamma".
- Repita o mesmo procedimento para a outra partição.



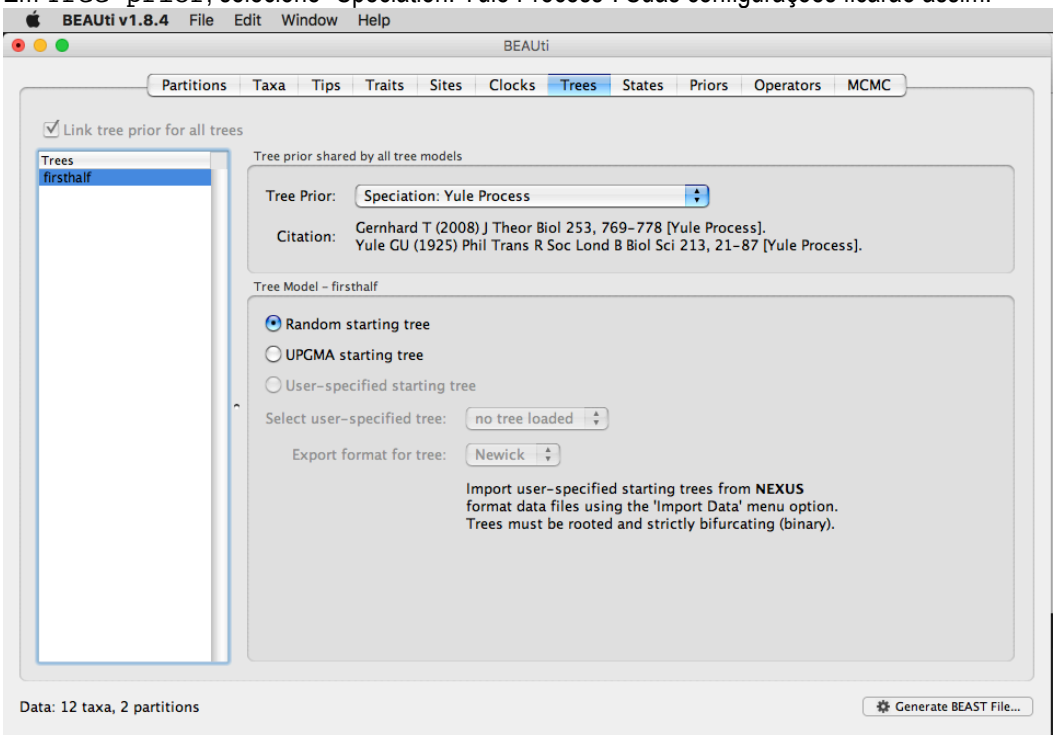
(6) Defina o modelo de relógio (*clock model*).

- Agora, selecione a aba `Clocks` no topo da janela principal.
- Altere o `Clock type` para "Uncorrelated relaxed clock".
- Em seguida, aparecerá um novo item que pode ser alterado, `Relaxed distribution`; deixe-o em "Lognormal". Suas configurações deverão ficar dessa maneira:



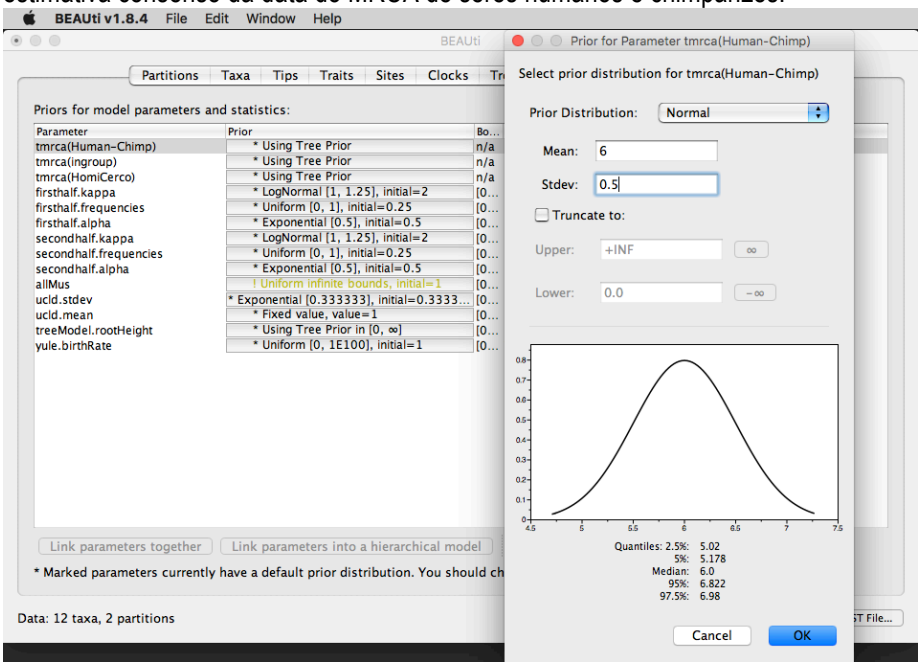
(7) Defina os *Tree priors.*

- Agora, selecione a aba `Trees` no topo da janela principal.
- Em `Tree prior`, selecione "Speciation: Yule Process". Suas configurações ficarão assim:

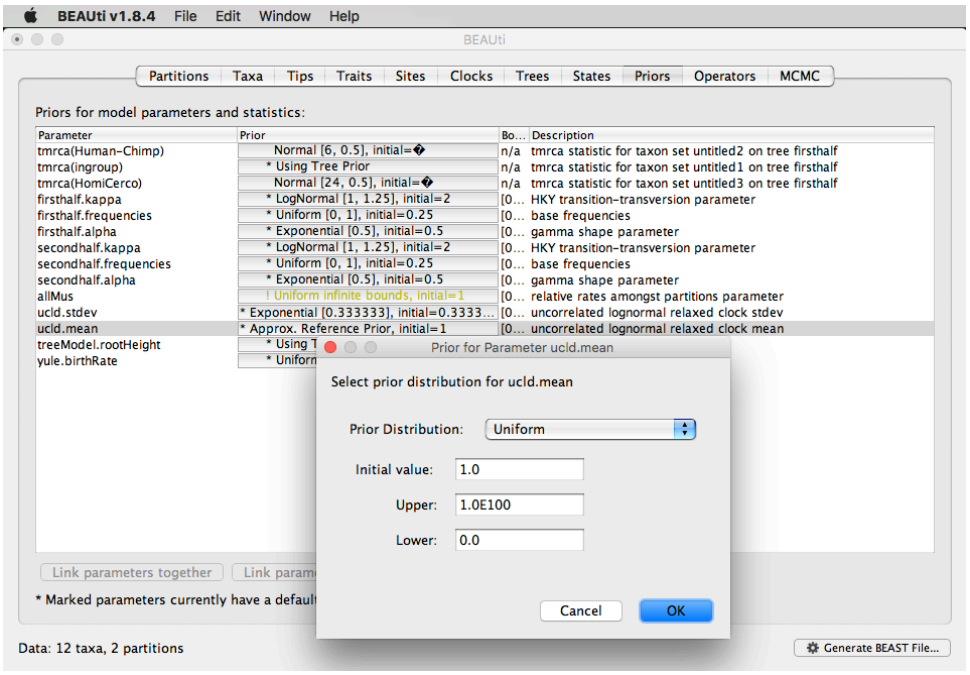


(8) Defina os *Priors.*

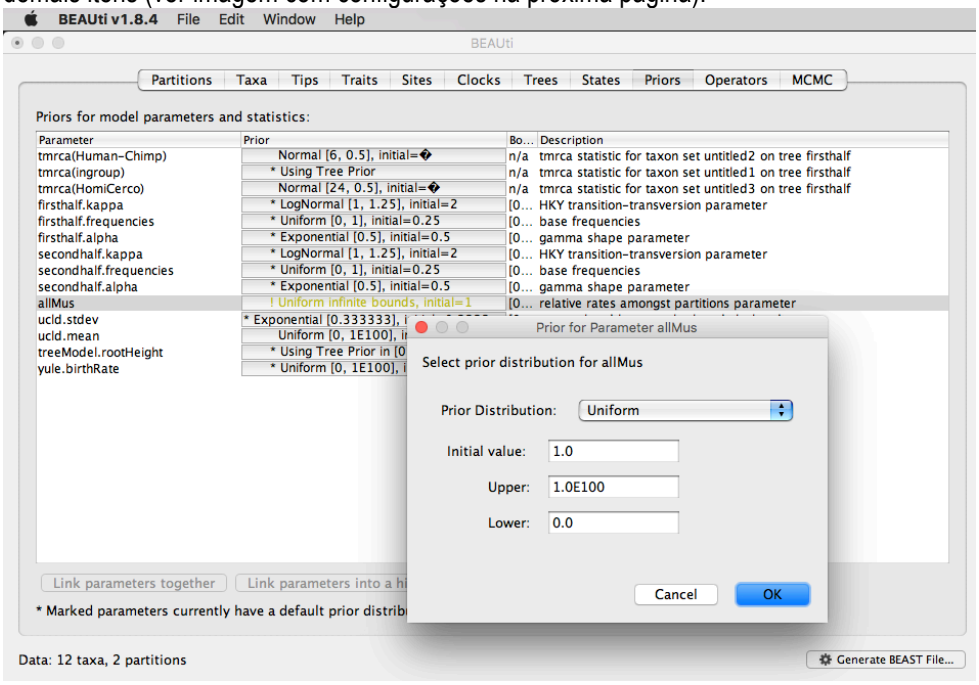
- Selecione a aba `Priors`, na qual especificaremos as *prior distributions* para os tempos de divergência baseado em conhecimento proveniente dos fósseis. Isso é chamado de "calibração da árvore". Nesse caso, faremos duas calibrações.
- Na coluna `Prior`, clique no botão referente ao parâmetro "tmrca(Human-Chimp)"; uma nova janela aparecerá, permitindo que você especifique o *prior* para esse parâmetro.
- Em `Prior Distribution`, selecione "Normal" e novos atributos ficarão disponíveis para alteração.
- Assumiremos uma distribuição normal centrada em 6 milhões de anos (`Mean`) com um desvio padrão (`Stdev`, *standart deviation*) de 0.5 milhão de ano. Isso resultará em um intervalo central de 95% de ca. 5-7 milhões de anos, correspondendo à estimativa consenso da data do MRCA de seres humanos e chimpanzés.



- Seguindo o mesmo procedimento, para tmrca(HomiCerco), defina a calibração de `Mean` 24 e +/- 0.5 `Stdev`.
- Na coluna `Prior`, clique no botão do parâmetro "ucld.mean" (referente ao modelo de relógio); em `Prior Distribution`, selecione "Uniform", e não altere os demais itens.



- Na coluna `Prior`, clique no botão do parâmetro "allMus"; em `Prior Distribution`, selecione "Uniform", e não altere os demais itens (ver imagem com configurações na próxima página).

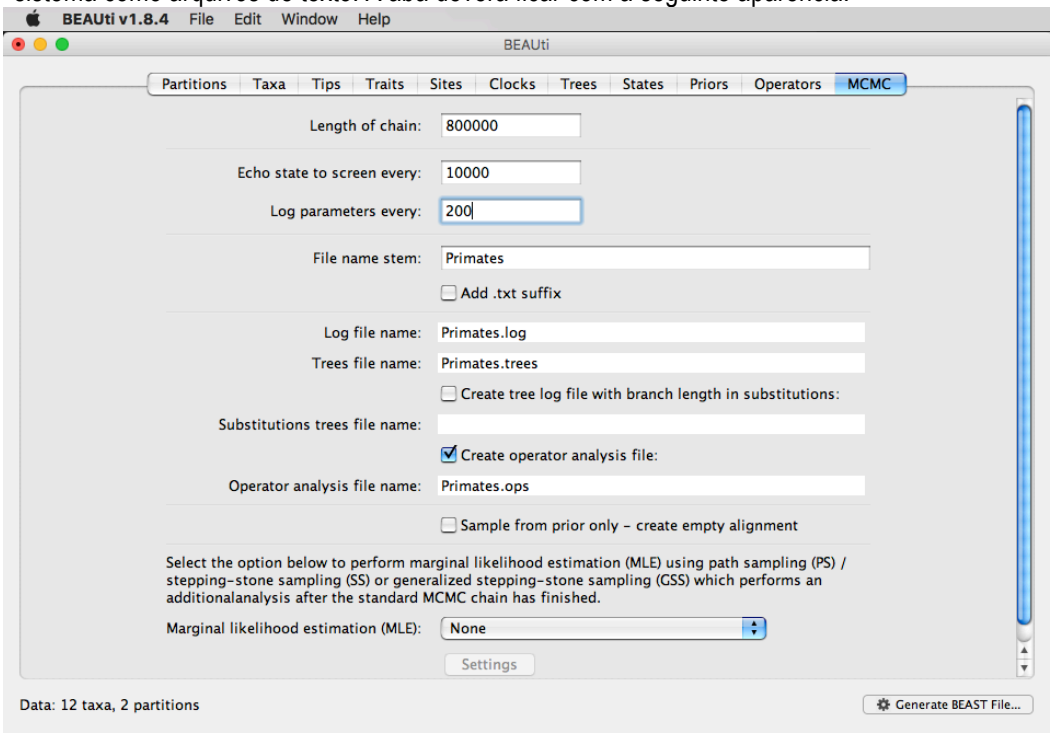


(9) Defina as opções para MCMC.

- Selecione a aba `MCMC`; é nessa aba em que especificaremos as opções que controlam a duração da análise e nomes dos arquivos.
- O item comprimento da cadeia (`Length of chain`) estabelece o número de gerações que a MCMC fará na cadeia. O valor *default* 10.000.000 é completamente arbitrário e deve ser alterado de acordo com o tamanho do conjunto de dados e a

complexidade do modelo. Para esses dados, vamos definir o comprimento de 800.000; isso deve resultar em uma corrida rápida (poucos minutos) na maioria dos computadores modernos.

- O próximo item especifica a frequência com que os valores dos parâmetros da cadeia de Markov serão exibidos em tela e gravados no *log file*. As informações que aparecem em tela têm a função de permitir o monitoramento do progresso do programa, então podem ser definidos como qualquer valor (embora se for definido um valor muito baixo, a análise na verdade ficará mais lenta). Para essa análise, coloque em `Echo state to screen every` o valor 10.000.
- Para o *log file*, o valor deve ser definido em relação ao comprimento total da cadeia. Uma amostragem muito frequente resultará em arquivos muito grandes com pouco benefício em termos de precisão da análise. Se a amostragem for pouco frequente, o *log file* não conterá muita informação sobre a distribuição dos parâmetros. Seu objetivo deve ser armazenar não mais do que 10.000 amostras. Para esse exercício, coloque em `Log parameters every` o valor 200.
- Os dois próximos itens nomeiam os *log files* dos parâmetros amostrados e das árvores. Esses serão definidos automaticamente com base no nome do arquivo NEXUS que foi importado (mas você pode alterar se quiser). Se você estiver usando um computador com Windows, sugerimos que você adicione o sufixo .txt aos arquivos (selecione o *checkbox* "Add .txt suffix"), de modo que os arquivos sejam nomeados como Primates.log.txt e Primates.trees.txt sejam reconhecidos pelo sistema como arquivos de texto. A aba deverá ficar com a seguinte aparência:

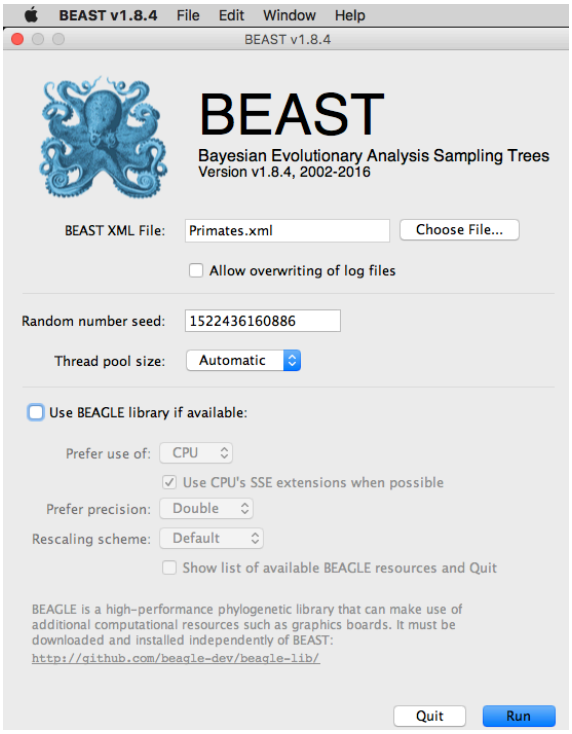


(10) Gere o arquivo XML para BEAST.

- Para criar o arquivo XML para o BEAST, clique no botão `Generate BEAST File...` no canto inferior direito da janela (ou selecione essa opção pelo menu `File`).
- Uma nova tela se abrirá. Verifique os *priors* que ficaram com os atributos *default* e clique no botão `Continue`.
- Salve o arquivo com um nome apropriado, usando .xml como extensão. Nesse caso, nomeie o arquivo como Primates.xml.
- Agora, estamos prontos para correr a análise no BEAST.

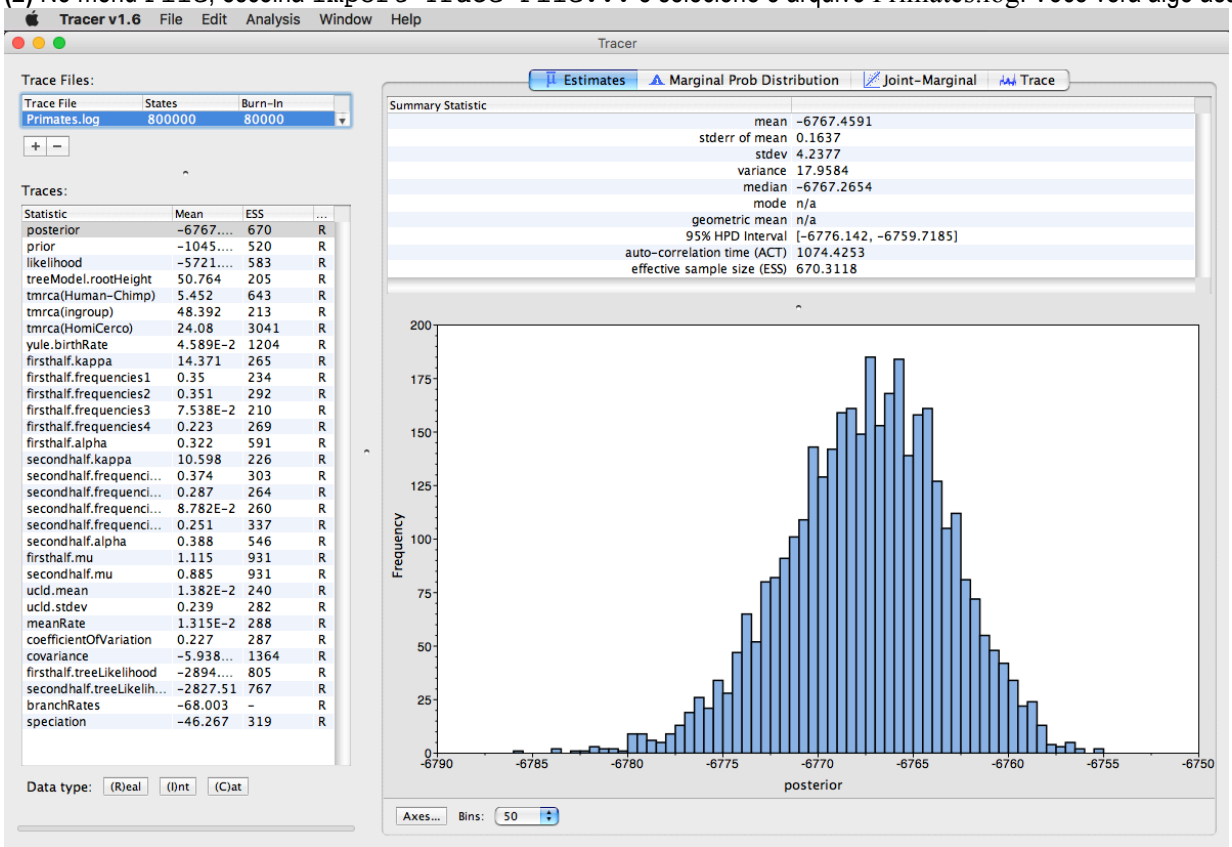
PASSO 2: BEAST

(1) Execute o programa BEAST.
(2) Clique no botão Choose File... para usar como *input* o arquivo XML que você acabou de criar.
(3) Se `Use Beagle library if available` estiver selecionado, desmarque a seleção. A tela deverá ficar com a aparência da figura ao lado.
(4) Atenção, os resultados serão salvos no mesmo lugar onde estiver seu *input file*.
(5) Clique no botão `Run` e a análise será iniciada. O progresso da análise será exibido na tela. A análise estará finalizada quando o BEAST parar de exibir novas informações na tela (e deve aparecer o tempo total de duração da análise).
(6) Verifique que os arquivos Primates.log e Primates.trees foram salvos no local em que está o seu *input file*.



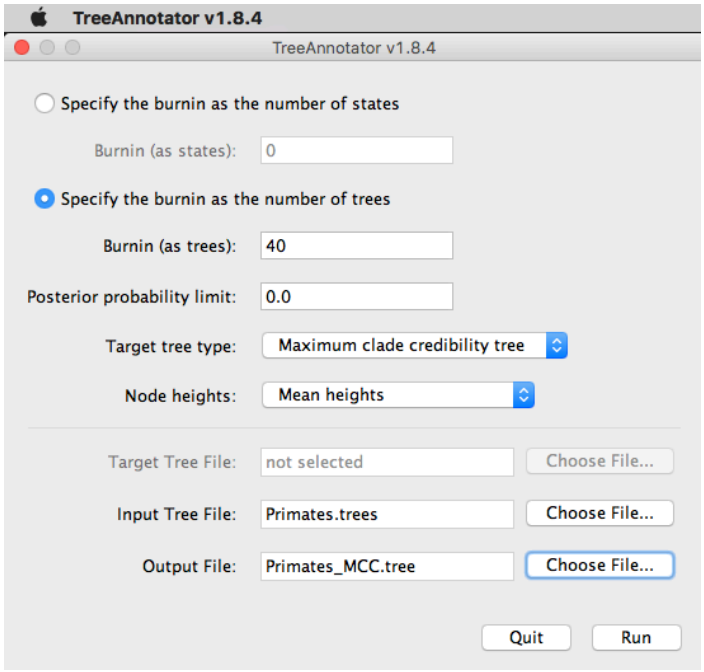
PASSO 3: Tracer

(1) Execute o programa Tracer para analisar o *output* do BEAST.
(2) No menu `File`, escolha `Import Trace File...` e selecione o arquivo Primates.log. Você verá algo assim:



(3) Lembre-se de que a MCMC é um algoritmo estocástico, então os números não serão exatamente os mesmos.
(4) À esquerda há uma lista com os registros do BEAST. Existem traços (`Traces`) para a distribuição posterior (esse é o log do produto da probabilidade da árvore e das probabilidades dos *priors*) e de parâmetros contínuos. Ao selecionar um traço à esquerda, seus valores serão evidenciados à direita, de acordo com a aba que estiver selecionada (ao lado direito, selecionada na parte superior).
(5) Quando o arquivo é aberto pela primeira vez, o traço *posterior* estará selecionado e as diversas estatísticas desse traço serão exibidas na aba `Estimates`.
(6) O objetivo dessa atividade é apenas o manuseio desses programas, então, não nos aprofundaremos nos significados dessas estatísticas.

PASSO 4: TreeAnnotator
(1) Um dos *outputs* do BEAST é uma amostragem de árvores plausíveis juntamente com seus parâmetros estimados que precisam ser sumarizados usando um programa chamado TreeAnnotator. Esse programa pegará esse conjunto de árvores e encontrará a com melhor sustentação. Em seguida, o programa anotará essa árvore resumo (*summary tree*) com as idades médias de todos os nós, os intervalos de confiança (*highest probability density*, *HPD*) e probabilidade posterior para cada nó.
(2) Para fazer tudo isso, execute o programa TreeAnnotator.
(3) O *burnin* é o número de árvores (ou gerações) a serem removidas do início da amostragem. Para essa análise, especificamos 800.000 gerações para comprimento da cadeia, amostrando a cada 200 gerações, então, esse arquivo de árvores produzido pelo BEAST contém 4000 árvores. No item `Specify the burnin as the number of trees`, para especificar um `Burnin (as trees)` de 1%, use o valor 40.
(4) O item `Posterior probability limit` especifica um limite de modo que se um nó encontra-se abaixo da frequência indicada na amostra de árvores (isto é, tem uma probabilidade posterior inferior a esse limite), ele não será anotada. Deixe esse valor em 0.0 para que todos os nós sejam anotados.
(5) No item `Target tree type` escolha a opção "Maximum credibility tree", para o programa encontrar a árvore com o maior produto da probabilidade posterior de todos os seus nós.
(6) Em `Node heights`, selecione "Mean heights" para que a altura (idade) de cada nó na árvore seja definida de acordo com a altura média ao longo de toda a amostra de árvores.
(7) Em `Input Tree File`, selecione o arquivo criado pelo BEAST Primates.trees e escolha um nome para o output do TreeAnnnotator (sugiro Primates_MCC.tree). A configuração completa ficará com a seguinte aparência:



(8) Pressione o botão Run e aguarde a mensagem na tela indicando a finalização da anotação.
(9) O arquivo Primates_MCC.tree será salvo no local em que está o seu *input file*.

PASSO 5: FigTree

(1) E, finalmente, podemos examinar a árvore. Para tanto, execute programa FigTree.

(2) No menu `File`, escolha a opção `Open` para abrir o arquivo Primates_MCC.tree.

(3) No painel de controle à esquerda, você pode selecionar o que deseja visualizar. Selecione o *checkbox* de `Node bars` e clique sobre a seta da esquerda para abrir as opções. Em `Display`, selecione a opção "height_95%_HPD", para visualizar as barras de erros das estimativas de idade.

(4) Selecione o *checkbox* de `Branch labels` e clique sobre a seta da esquerda para abrir as opções. Em `Display`, selecione a opção "posterior", para visualizar a probabilidade posterior de cada nó. A imagem deve parecer com a figura abaixo: