

Práctica 4: El misterioso homínido de Siberia

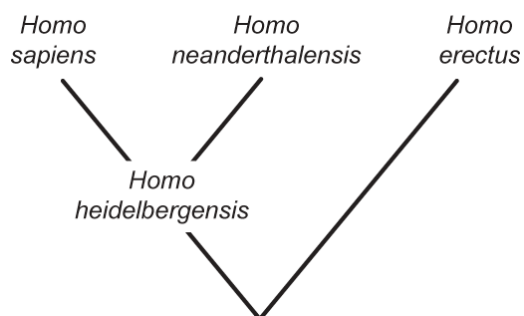
Introducción

En 2008, la falange del dedo de un homínido no identificado fue excavada de la Cueva Denisova, en las montañas Altai de Siberia, en Rusia. La cueva ya era conocida por tener evidencia de habitación humana remontándose >125000 años al pasado.

Usando estimativos de fechas de descomposición del radiocarbono, la edad del nuevo espécimen fue estimada entre 30000 y 48000 años. Esto significa que el individuo existió en tiempos en los que los humanos modernos (*Homo sapiens*) y los neandertales (*Homo neanderthalensis*) cohabitaron en extendidas regiones de Eurasia, antes de la extinción de los neandertales hace aproximadamente 25000 años.



Antes de la aparición de los humanos modernos y los neandertales, hubo otra especie humana que estaba presente en Eurasia: *Homo erectus*. Hay evidencia que *Homo erectus* migró desde África hace alrededor de 1.9 millones de años, posiblemente existiendo en Indonesia hasta hace 100000 años. Esto es diferente a la historia de los humanos modernos, quienes salieron de África hace ~50000 años a colonizar Eurasia, de acuerdo a evidencia genética y arqueológica.



La opinión actual de la filogenia de los homínidos es que los humanos modernos y los neandertales son especies hermanas, posiblemente compartiendo un ancestro a *Homo heidelbergensis*, hace aproximadamente medio millón de años. *Homo erectus* se considera un pariente más lejano. Hace aproximadamente 6 o 7 millones de años, el linaje que llevó a todas estas especies de *Homo* divergió del linaje que llevó a las dos especies actuales de chimpancé (*Pan troglodytes* y *Pan paniscus*). Juntas, estas especies forman un grupo llamado "Hominini".

Investigadores del Instituto Max Planck de Antropología Evolutiva en Leipzig secuenciaron el genoma mitocondrial de la falange de Denisova. Secuenciar ADN antiguos de homínidos es un verdadero reto. El ADN está altamente degradado y con una baja concentración de ADN auténtico, lo que significa que hay un riesgo alto de contaminación de otras fuentes. Sin embargo, nuevas tecnologías de secuenciación han sido claves para sobrellevar estos retos.

En esta práctica, vas a investigar al homínido de Denisova usando un análisis filogenético Bayesiano. El análisis te permitirá elucidar las relaciones entre el misterioso individuo y otros homínidos, y estimar la escala temporal de estos eventos evolutivos.

Práctica 4 : Análisis filogenético Bayesiano de los homínidos

Antes de comenzar, confirma que tengas versiones recientes de los siguientes programas:

- *BEAST 2* (que incluye *BEAUti*, *BEAST*, y *TreeAnnotator*)
- *Tracer*
- *FigTree*

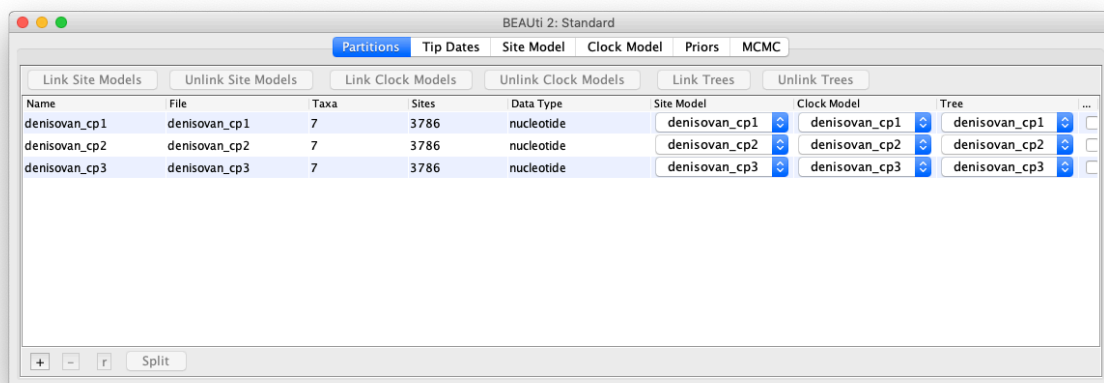
Los tres archivos de datos para esta sección de los ejercicios prácticos son **denisovan_cp1.nex**, **denisovan_cp2.nex**, y **denisovan_cp3.nex**. Cada archivo está en formato 'Nexus', que es ampliamente usado para filogenética. Los alineamientos son las posiciones 1, 2, y 3 de los codones de 13 genes mitocondriales codificantes, para 7 especies de homínidos: (i) el homínido de Denisova; (ii) Neandertal (*Homo neanderthalensis*); (iii) humano moderno (*Homo sapiens*); (iv) chimpancé común (*Pan paniscus*); (v) gorila (*Gorilla gorilla*); y (vi) orangután (*Pongo pygmaeus*).

En éste ejercicio usaremos el programa de filogenética Bayesiana *BEAST 2*. El programa es extenso y requiere que creemos un archivo en formato XML, que podemos crear usando el intuitivo programa *BEAUti*. Hay tres partes al análisis:

- Crear un archivo en formato XML usando *BEAUti*
- Análisis Bayesiano usando *BEAST*
- Procesar los resultados usando *Tracer*, *TreeAnnotator*, y *FigTree*

Sección 4.1: Crear un archivo en formato XML usando *BEAUti*

- Abre el programa *BEAUti*. El propósito de este programa es crear un archivo para *BEAST*. El primer paso es cargar nuestros alineamientos a *BEAUti*. Selecciona "Import Alignment" de el menú "File" (archivo) y abre los tres alineamientos, **denisovan_cp1.nex**, **denisovan_cp2.nex**, y **denisovan_cp3.nex**. Alternativamente, puedes arrastrar los archivos de los alineamientos a la ventana de *BEAUti*.
- Deberías estar en la pestaña **Partitions** en *BEAUti*. La ventana muestra algunas de las características de los datos que has cargado. Podrás ver que cada alineamiento contiene 7 organismos y 3786 nucleótidos.



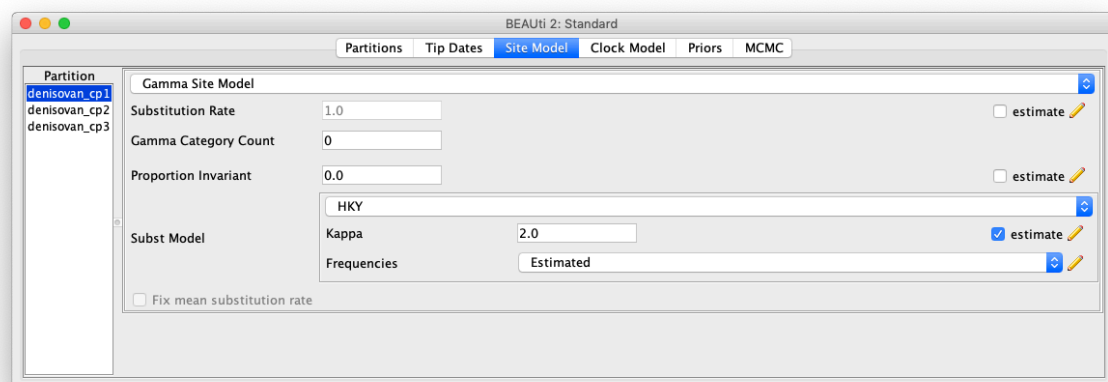
Hay varias opciones en esta pestaña relacionadas a cómo particionar los datos. Hacer particiones nos permite aplicar diferentes modelos evolutivos a cada parte de los alineamientos.

Por defecto, *BEAUti* le asigna a cada una de las tres posiciones de codón diferentes modelos de sustitución (site), tasa evolutiva entre especies (clock), y árbol. Esto se ve en el hecho que hay tres nombres diferentes en las filas de las columnas para estos modelos. Para este análisis queremos que cada posición tenga diferentes modelos de sustitución y tasas evolutivas. Sin embargo, podemos asumir que las tres posiciones comparten el árbol filogenético porque están ligadas en el genoma de la mitocondria, el cual no recombina.

Selecciona los tres alineamientos en esta pestaña y haz click en “Link Trees”. Puedes ver que ahora los árboles de las 3 posiciones de codón tienen el mismo nombre “denisovan_cp1”.

- c) Ignora por ahora la pestaña **Tip Dates**. Normalmente, esa pestaña nos permite incluir las edades del homínido de Denisova y el Neandertal para el análisis, pero esto no lo haremos para reducir la carga computacional. También ignoraremos este paso porque las edades de estas secuencias son mínimas comparado con la edad total de el árbol filogenético que estamos investigando.
- d) Dirígete a la pestaña **Site Model**, donde elegiremos el modelo de sustitución. En este análisis vamos a usar el modelo HKY para cada una de las tres posiciones de codón. El modelo HKY permite que la tasa de transiciones ($A \leftrightarrow G$ and $C \leftrightarrow T$) sea diferente a la tasa de transversiones ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$ and $G \leftrightarrow T$). Para seleccionar este modelo, selecciona “HKY” en la caja “Subst Model”. Deja el valor por defecto de 2.0 para el parámetro “Kappa”, que representa la proporción de transiciones contra transversiones. Este parámetro será estimado en el análisis. Deja también “Frequencies” como “Estimated”, que significa que vamos a estimar las frecuencias ancestrales de cada uno de los nucleótidos.

Ahora selecciona todas las particiones en la ventana a la izquierda de la pestaña. Ahora veras una opción “Clone from denisovan_cp1”, que significa que *BEAUti* nos esta dando la opción de copiar el modelo de sustitución para las otras 2 posiciones de codón. Haz click en “OK” para asignarles el modelo HKY a las otras 2 posiciones de codón. Ahora cada una de las 3 posiciones tendrá su propio modelo HKY.



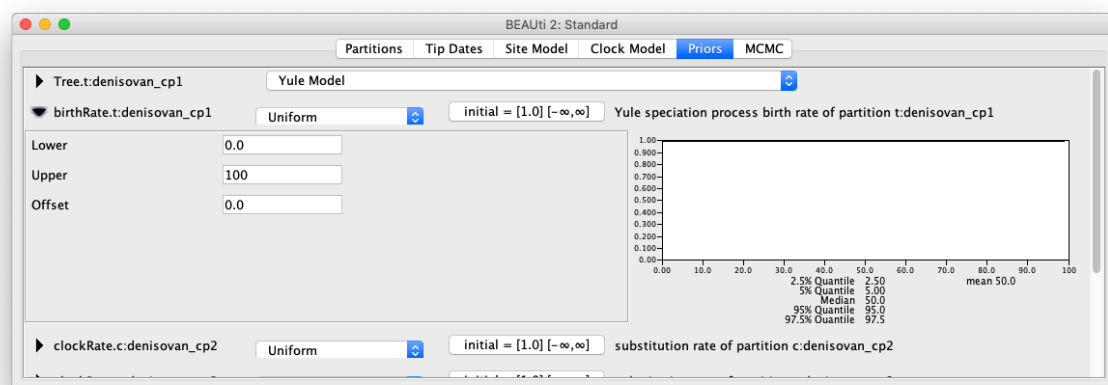
- e) Dirígete a la pestaña **Clock Model**. Aquí necesitamos elegir el tipo de modelo de tasas evolutivas entre especies (reloj molecular) que queremos usar en nuestro análisis. Fíjate que aunque no vamos a estimar fechas de divergencia en este análisis específico, *BEAST* exige que elijamos un modelo de reloj molecular. *BEAST* solo infiere árboles enraizados. Vamos a usar un reloj molecular estricto, “strict clock”, para cada una de las tres posiciones de codón. Puedes dejar el valor de “Clock.rate” en el defecto de 1.0.

El modelo que hemos elegido es un reloj molecular estricto para cada una de las tres posiciones de codón. Esto permite que cada posición de codón tenga su propia tasa evolutiva. Específicamente, el programa fijará la tasa evolutiva de la posición de codón 1 a 1.0, mientras las tasas de las otras dos posiciones serán estimadas relativas a la de la posición 1 durante el análisis.

- f) Dirígete a la pestaña **Priors**. Aquí tenemos que elegir distribuciones para los priors de cada uno de los parámetros de nuestro análisis, incluyendo el árbol. En el menú a la derecha de “Tree.t:denisovan_cp1” hay varios modelos que se pueden usar para generar una distribución para el árbol. En este análisis vamos a lidiar con secuencias de diferentes especies, que significa que necesitamos usar un modelo de especiación. Los modelos coalescentes solo son apropiados para datos a nivel poblacional. Elige el modelo de especiación más simple, que es el proceso Yule. Este modelo es de solo especiación, en el que todos los linajes tienen la misma probabilidad de bifurcarse en dos linajes descendientes, en todos los momentos de tiempo.

Si haces click en la flecha negra a la izquierda de “Tree.t:denisovan”, veras las opciones para el modelo Yule. Deja el valor inicial para la tasa de especiación (birth rate) en el defecto de 1.0.

Mira los otros priors de esta pestaña. Hay una distribución uniforme para la tasa de especiación del modelo Yule, y una distribución lognormal para el parámetro kappa para cada una de las tres posiciones de codón. Debemos cambiar la distribución uniforme de el prior de la tasa de especiación para que su limite superior no sea infinito. Haz click en el triangulo negro al lado de “birthrate.t:denisovan” y cambia el valor de “Upper” (límite superior) a 100.



- g) Dirígete a la pestaña **MCMC**. Aquí tenemos que definir cuanto tiempo vamos a pasar acumulando muestras de la distribución a posterior usando simulaciones en una cadena de Monte Carlo markoviana (MCMC). Recuerda que queremos estimar las distribuciones

a posterior de los parámetros, incluyendo el árbol. Sin embargo, esta distribución no se puede calcular directamente, sino que tenemos que tomar muestras de distribuciones a posterior usando una simulación MCMC que esté correctamente diseñada. Usando estas muestras Podemos aproximar la distribución a posterior. Para que el análisis sea rápido, cambia la longitud de la cadena, “Chain Length” a 5000000.

Haz click en el triangulo negro al lado de “tracelog” y cambia el nombre del archivo “File name” a **denisovan_hky.log**. Ahora haz click en el triangulo negro al lado de “treelog.t:denisovan_cp1” y cambia el nombre del archivo de árboles a **denisovan_hky.trees**.

- h) Ahora dirígete al menú File (o Archivo) y selecciona “Save As” (Guardar como). Guarda el archivo XML con el nombre **denisovan_hky.xml** en el escritorio de tu computador o en tu directorio actual (donde sea que quieres que *BEAST* escriba los resultados). Esto va a escribir un archivo en formato XML que *BEAST* puede usar. Mantén *BEAUi* abierto, ya que luego vamos a cambiar algunas de las especificaciones del análisis.

Sección 4.2: Análisis filogenético Bayesiano usando *BEAST*

- a) Abre el programa *BEAST* y elige el archivo XML que creaste en la sección previa. Haz click en el botón “Run” en la parte inferior derecha de la ventana.
- b) Mientras el análisis progresa, *BEAST* continuamente escribirá los resultados en tres archivos nuevos. El que termina en .log contiene los valores de la distribución a posterior de los parámetros de los modelos, mientras el archivo terminado en .trees contiene las muestras de la distribución a posterior de árboles. El archivo terminado en .state guarda el estado actual de la cadena MCMC, en caso que quieras resumir el análisis desde un momento dado.

El análisis se tomará unos minutos dependiendo de la velocidad de tu computador. Mientras esperas, puedes probar responder estas preguntas sobre la filogenética Bayesiana.

Q. *¿Cómo podemos elegir la distribución prior para cada parámetro en el análisis?*

.....
.....
.....
.....

Q. *¿Sería correcto usar estimativos de nuestros datos para informar nuestra elección de la distribución para el prior? ¿Por qué o por qué no?*

.....
.....

.....

.....

Q. *¿Dado que no es posible obtener la distribución a posterior directamente, qué método podemos usar para estimar la distribución a posterior?*

.....

.....

.....

Sección 4.3: Procesar los resultados usando *Tracer*, *TreeAnnotator*, y *FigTree*

- a) Abre el programa *Tracer*, y haz click en “Import Trace File” en el menú “File” para importar el archivos .log files de tu análisis de *BEAST*.
- b) Puedes inspeccionar las características de las distribuciones a posterior para cada parámetro. Lo primero es mirar que el número efectivo de muestras (ESS; effective simple size) sea superior a 200 para todos los parámetros. Esto indica que se han tomado suficientes muestras de la distribución a posterior para cada parámetro. Las muestras del MCMC no son completamente independientes entre ellas, y esta es la razón que en número efectivo de muestras es menor al total. Si algún valor de ESS es inferior a 200, indica que es necesario tomar más muestras durante el análisis. Si esto ocurre en tus parámetros, lo puedes ignorar para el proposito de esta práctica.
- c) En adición, queremos que las muestras vengan de la distribución estacionaria. Por esta razón normalmente nos deshacemos de el primer 10% de las muestras. Esto se conoce como la fase ‘burn-in’. Por defecto, *Tracer* excluye el primer 10% de tus muestras cuando calcula el promedio y otras estadísticas. Normalmente llevamos a cavo el análisis múltiples veces y revisamos que las replicas de los análisis tengan resultados consistentes. Puedes intentar hacer esto y revisar los resultados de múltiples análisis simultáneamente en *Tracer*.
- d) Miremos los estimativos de algunos de los parámetros. Específicamente, mira los parámetros “clockRate.2” y “clockRate.3”. Estos representan las tasas de sustitución para los codones 2 y 3, respectivamente, comparadas con el codón 1 (cuya tasa ha sido fijada en 1.0).
- Q. *¿Cuales son las tasas de sustitución relativas en las posiciones de codón 2 y 3? ¿Es este resultado consistente con nuestra expectativa del proceso de evolución molecular, dado nuestro conocimiento sobre los codones y cómo codifican amino ácidos?*

.....

.....

-
-
- e) Abre el programa *TreeAnnotator*. Este programa se usa para procesar el archivo terminado en .trees producido por *BEAST*. El programa leerá todas las muestras de árboles del MCMC y resumirá la información en un solo árbol.
 - f) Escribe el valor “10” en la caja al lado de “Burnin percentage”. Esto significa que estamos deshaciéndonos del primer 10% de las muestras, tratándolas como burn-in. En la opción “Input Tree File” haz click en “Choose File” y selecciona el archivo terminado en .trees generado por el análisis de *BEAST*. En la opción “Output File” haz click en “Choose File” y selecciona el directorio donde quieres guardar el archivo que va a producir *TreeAnnotator*. Dale el nombre **denisovan_hky.tre** y haz click en “Run”.
 - g) Abre el programa *FigTree* y úsalo para ver el archivo **denisovan_hky.tre** que acaba de generar *TreeAnnotator*. En la ventana verás el árbol que resume las muestras de arboles de tu análisis filogenético Bayesiano. Puedes explorar las opciones de *FigTree* para ver la información asociada con el árbol. Prueba hacer click en los tres símbolos en el menú “Layout”. De izquierda a derecho, estos deberían permitirte visualizar el árbol como enraizado, circular, y desenraizado, respectivamente.

Estamos particularmente interesados en dos características del árbol. Primero, queremos ver donde esta posicionado el homínido de Denisova. Haz click en la caja al lado de “Node Labels” y selecciona “posterior” en el menú al lado de “Display”. Esto pondrá información en los nodos del árbol de acuerdo a sus probabilidades a posterior, que indican el soporte estadístico para cada una de las agrupaciones representadas en el árbol.

- Q.** *¿Donde esta localizado el homínido de Denisova en el árbol filogenético? ¿Cual es la probabilidad a posterior de la agrupación de el homínido de Denisova junto con los otros dos humanos?*

.....

.....

.....

- Q.** *¿De acuerdo a este resultado, el homínido de Denisova es un humano moderno, un neandertal, o ninguno de los dos?*

.....

.....

.....

.....

Q. *Cuando tomas muestras durante el análisis usando MCMC, elegimos guardar el valor de los parámetros cada 1000 pasos. ¿Por qué no elegimos guardar los valores en todos y cada uno de los pasos?*

.....

.....

.....

.....

.....

Q. *¿Qué podemos hacer para reducir la duración del ‘burn-in’? En otras palabras, ¿qué podemos hacer para ayudar a nuestra cadena markoviana para que llegue a la distribución estacionaria más rápidamente?*

.....

.....

.....

.....

.....

Q. *A veces la cadena markoviana no explora en detalle el paisaje de posibles combinaciones de parámetros, de manera que nuestras muestras no proveen una buena representación de la distribución a posterior. ¿Cómo podemos identificar si este es el caso?*

.....

.....

.....

.....

.....

Práctica opcional: Estimación Bayesiana de tasas evolutivas

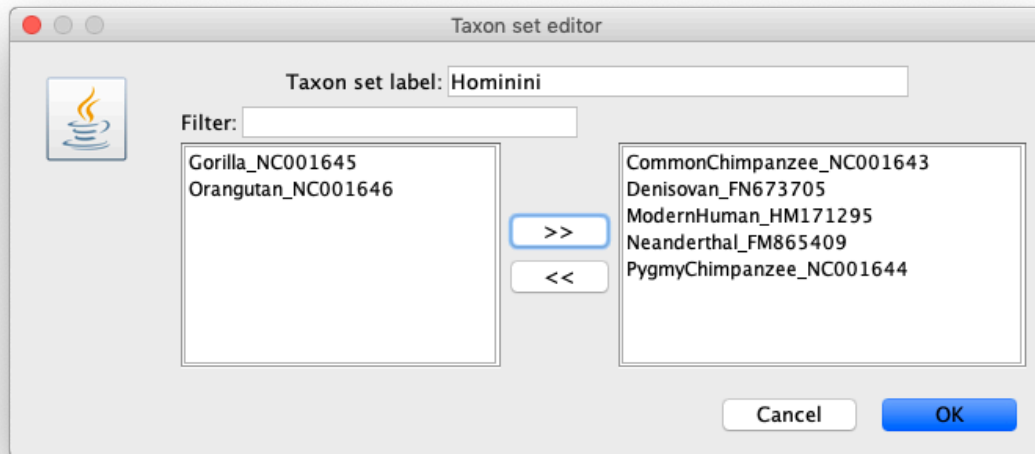
En esta sección vas a llevar a cabo más análisis de los datos de los homínidos para estimar sus tasas evolutivas y fechas de divergencia. Hay dos partes en el análisis:

- Análisis Bayesiano usando *BEAST* y un reloj molecular relajado
- Procesar los resultados usando *Tracer*, *TreeAnnotator*, y *FigTree*

Crear un archivo en formato XML usando *BEAUti*

- Abre el programa *BEAUti*. Selecciona “Import Alignment” del menú “File” y elige el archivo del alineamiento **denisovan.nex**. También puedes arrastrar los datos a la ventana de *BEAUti*. Este archivo contiene 13 genes mitocondriales codificantes unidos en un alineamiento (concatenados), correspondientes a los mismos 7 homínidos que exploramos en la sección A. Para esta sección práctica, vamos a simplificar las cosas al no usar una partición de los datos (como en la sección A). En otras palabras, vamos a usar un único modelo de sustituciones para modelar la evolución de estos homínidos.
- Dirígete a la pestaña **Site Model**. Selecciona el modelo de sustitución HKY+G con 4 categorías de la distribución gamma para las tasas de evolución de los sitios.
- Dirígete a la pestaña **Clock Model**. Selecciona “Relaxed Clock Log Normal”. Esto aplicará un modelo lognormal de reloj molecular relajado, que asume que las tasas en las ramas vienen de una sola distribución lognormal.
- Ve a la pestaña **Priors** y especifica un prior uniforme entre 10^{-6} a 1 sustituciones/sitio/Ma para el prior de la tasa promedio (“uclMean.c:denisovan”).
- También en la pestaña **Priors**. Selecciona un modelo Yule (“Yule Model”) para el prior del árbol y cambia el prior uniforme para la tasa de especiación para que tenga un rango entre 0 y 100.
- Ahora queremos definir grupos de especies que puedan ser de interés. En este análisis, estamos interesados en uno de los nodos del árbol que se pueda usar para una calibración de fecha. Podemos definir un grupo de organismos haciendo click en el botón “+ Add Prior” y adicionando un prior a la edad de un ancestro mas reciente (“MRCA prior”; most recent common ancestor prior).

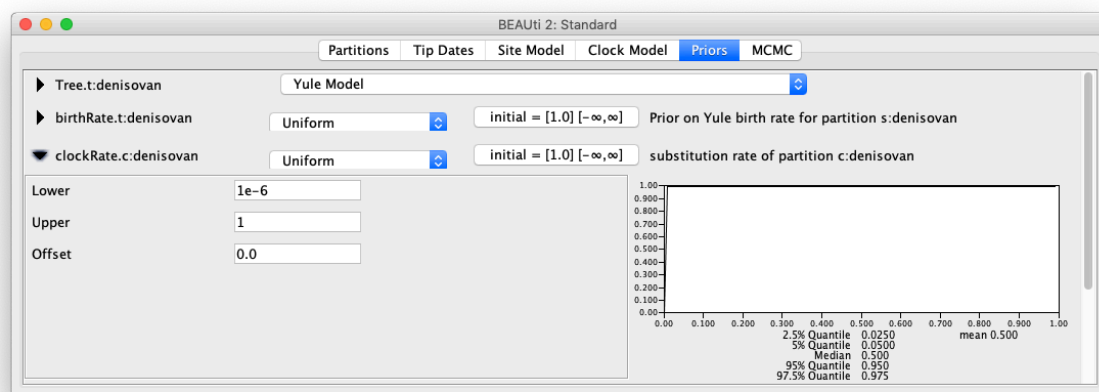
Para el nombre de este set de especies (“Label”), escribe “Hominini”. En este set de especies, queremos incluir tres humanos (el humano moderno, el neandertal, y el de Denisova), y dos chimpancés (chimpancé común, y el chimpancé pigmeo). Selecciona a estos grupos y haz click en las flechas hacia la derecha para incluirlos en el grupo de organismos. Luego haz click en “OK”.



- g) Ahora queremos especificar la distribución del prior para la edad de este nuevo grupo de organismos. Creemos que el ancestro común de los chimpancés y los humanos existió hace al rededor de 6.5 millones de años, y muy probablemente entre 6 y 7 millones de años. Podemos usar esta información para elegir los parámetros de una distribución normal. En el menú a la derecha de “Hominini.prior” selecciona “Normal”.

Ahora necesitamos definir los parámetros de nuestra distribución normal para el prior. Haz click en el triángulo negro a la izquierda y cambia los parámetros de la distribución normal para que el promedio (mean) sea de 6.5, y la desviación estándar (sigma) sea de 0.2551. Fíjate que estamos dando fechas en millones de años. Tomate un tiempo para mirar la distribución y sus características.

- h) El siguiente paso es definir los priors para la tasa de sustitución. Publicaciones previas sugieren que la tasa de evolución mitocondrial en animales cae en el rango entre 10^{-6} a 1 sustitución/sitio/Ma. Podemos usar estos valores para definir una distribución uniforme para la tasa de sustitución. Fíjate que el valor de 10^{-6} se escribe “1e-6”, como se en la figura siguiente.



- i) Dirígete a la pestaña **MCMC**. Para mantener la duración del análisis manejable, elige una longitud de cadena “Chain Length” de 10,000,000 pasos. Elige como nombre del archivo del rastro (trace) **denisovan_relaxedclock.log** y para el archive de árboles **denisovan_relaxedclock.trees**.
- j) Guarda todas tus especificaciones de *BEAST* en un archivo que puedes llamar **denisovan_relaxedclock.xml**. Mantén *BEAUti* abierto ya que te será útil para responder algunas de las preguntas abajo que se tratan de tus especificaciones.
- k) Lleva acabo un análisis de *BEAST* usando el nuevo archive XML. Si te parece que se va tardar demasiado tiempo, puedes detener y cerrar el análisis y trabajar con los resultados previamente completados para tu conveniencia (“pre-cooked runs”).

Processing the output using *Tracer*, *TreeAnnotator*, and *FigTree*

- a) Abre el programa *Tracer*. Haz click en “Import Trace File” en el menú “File” (Archivo), y carga el archivo terminado en .log de tu análisis de *BEAST*.
- h) Verifica que el número efectivo de muestras (effective sample size; ESS) de todos los parámetros muestreados sea mayor a 200. Esto indica que se han tomado suficientes muestras de la distribución a posterior para cada parámetro. El ESS es menor a el número total de muestras porque las muestras usando MCMC no son completamente independientes. Si algún valor de ESS es inferior a 200, indica que es necesario tomar más muestras durante el análisis. Si esto ocurre en tus parámetros, lo puedes ignorar en el caso de esta práctica. También puedes probar cambiar la cantidad de tus muestras que consideras burn-in, y ver si tu ESS incrementa.
- b) Mira el estimativo del “coefficientOfVariation”. Este es el coeficiente de variación en las tasas evolutivas de las ramas, calculado como la desviación estándar de las tasas dividida por el promedio. Esto provee una medida de la variación de la tasa en las ramas del árbol, donde un valor de 0 indica un reloj molecular estricto (cero variación entre ramas).

Q. *¿Ves que la distribución posterior del coeficiente de variación (de tasas en las ramas) incluya cero? Si no, ¿qué nos indica esto sobre la cantidad de variación en las tasas de las ramas?*

.....

.....

.....

.....

- c) Abre el programa *TreeAnnotator*. En la caja al lado de “Burnin percentage”, escribe “10”. Esto significa que estamos deshaciéndonos de el primer 10% de nuestras muestras como burn-in. Define en la caja “Input Tree File” el archivo terminado en .trees producido por

BEAST durante el análisis. En la caja “Output File” selecciona el directorio donde quieres que el programa *TreeAnnotator* guarde su resultado. Nombra el archivo que se va a producir **denisovan_relaxedclock.tre** y haz click en “Run”.

- d) Abre el programa *FigTree* y úsalo para ver el archivo **denisovan_relaxedclock.tre** que produjo *TreeAnnotator* en el paso anterior.

Aquí estamos interesados en la escala de tiempo evolutivo. En la caja “Node Labels” selecciona “height” en el menú que esta al lado de “Display”. Esto le pondrá información a los nodos del árbol con sus fechas estimadas en millones de años. También puedes ver los intervalos del rango de edades con 95% de la densidad a posterior más alta (highest posterior density; HPD), usando el menú al lado de “Display”.

- Q. *¿Cual es el promedio y el intervalo 95% HPD para el estimativo de la edad de la divergencia entre el homínido de Denisova y los otros dos humanos?*

.....
.....

- e) En la sección “Appearance” dales color a las ramas de acuerdo a su tasa evolutiva (“rate”). Explora el gradiente de colores para mostrar las tasas bajas en azul y las altas en rojo.

- Q. *¿Cuales ramas del árbol parecen tener una tasa evolutiva alta?*

.....
.....

- Q. *El análisis de divergencia lo calibramos usando una distribución normal para la edad de Hominini. Sin embargo, la distribución lognormal suele ser considerada la distribución paramétrica más adecuada para representar las edades de fósiles. ¿Por qué crees que este es el caso?*

.....
.....
.....
.....
.....
.....
.....

.....

Q. *¿Cómo opinas que podríamos mejorar la precisión de estos estimativos de fechas de divergencia moleculares?*

.....

.....

.....

.....

.....

.....

.....

.....