

# Análisis Filogenético de Brotes utilizando información genética

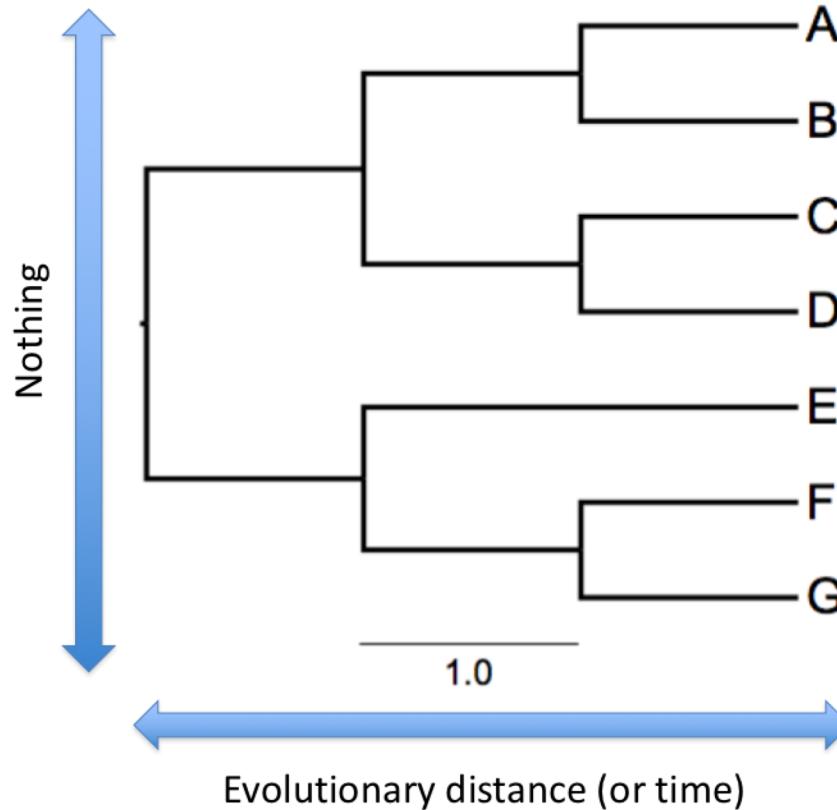
## Análisis Filogénetico Bayesiano con BEAST

Outbreak Analysis and Modelling for Public Health  
17-21 June 2019. Bogotá D.C - Colombia

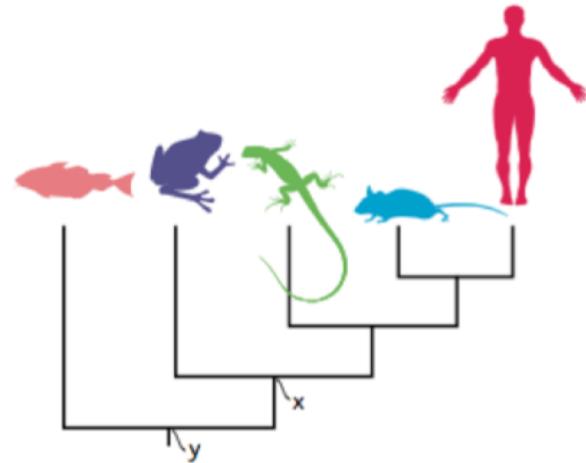
**Igor Siveroni**

MRC Centre for Global Infectious Disease Analysis  
Department of Infectious Disease Epidemiology  
Imperial College London

- Filogénetica  
*Likelihood filogenética / Algoritmo de Felsenstein*
- BEAST  
Bayesian Evolutionary Analysis by Sampling Trees
- Filodinámica
- PhyDyn  
Inferencia Filodinámica Bayesiana con Modelos Complejos  
*Likelihood Coalescente Estructurado*

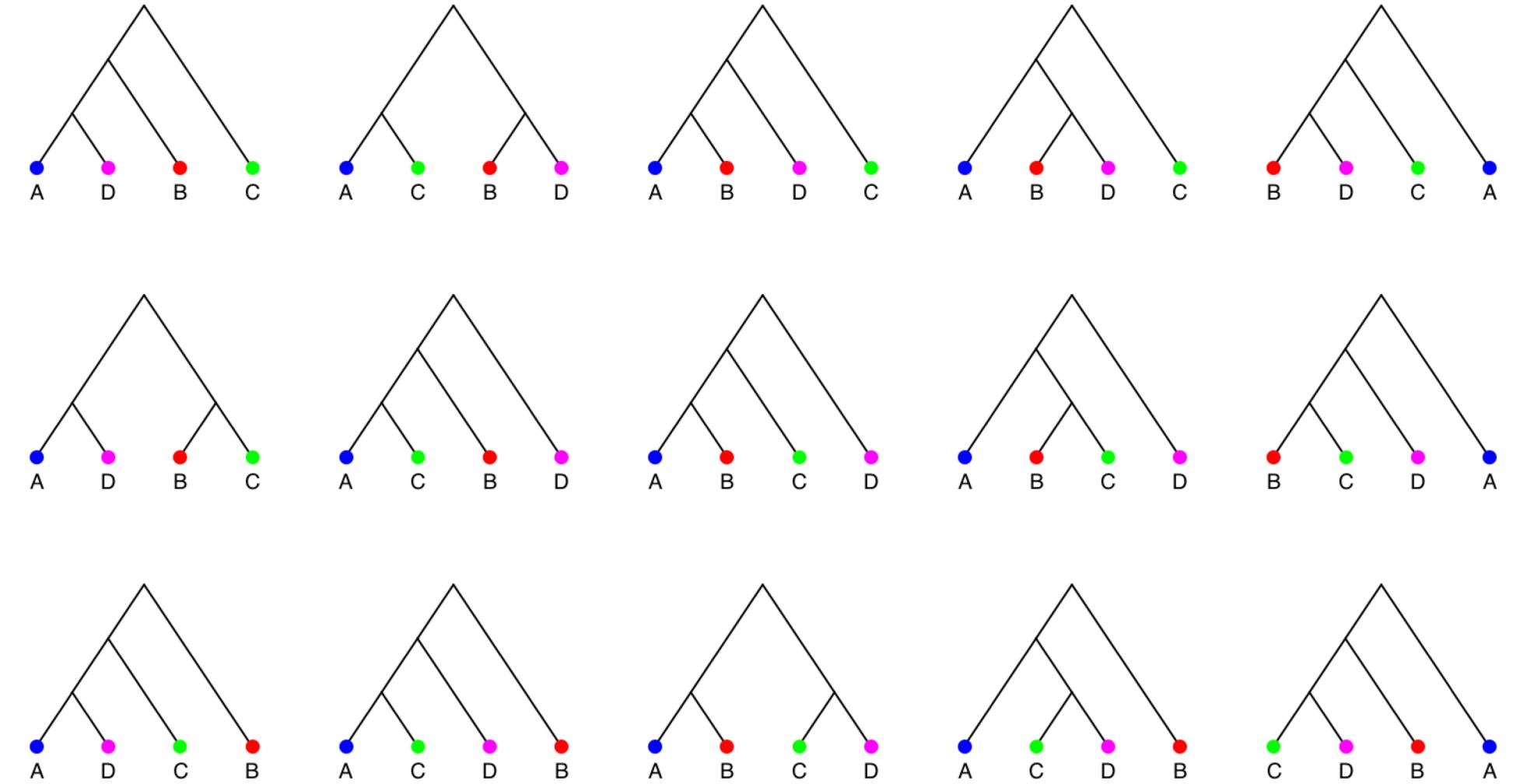


- La genealogia es el registro de reproduccion de una poblacion.
- La filogenia es un resumen, expresado como un arbol de diversidad genetica o fenotipica observada
  - Basado en un muestreo de la poblacion
  - Puede coincidir con la genealogia



Is the frog more closely related to the fish or the human?  
**The frog is more closely related to the human.**

¿Cuantos árboles existen  
para  $n$  taxa?



15 arboles posibles (sin rango) utilizando 4 individuos/especies



105 arboles posibles (sin rango) utilizando 5 individuos/especies



945 arboles posibles (sin rango) utilizando 6 individuos/especies

Para  $n$  species existen

$$T_n = 1 \times 3 \times 5 \times \cdots \times (2n - 3) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$$

árboles binarios etiquetados con raíz (rooted)

**Table 3.1 The numbers of unrooted trees ( $U_n$ ), rooted trees ( $R_n$ ) and labelled histories ( $H_n$ ) for  $n$  species**

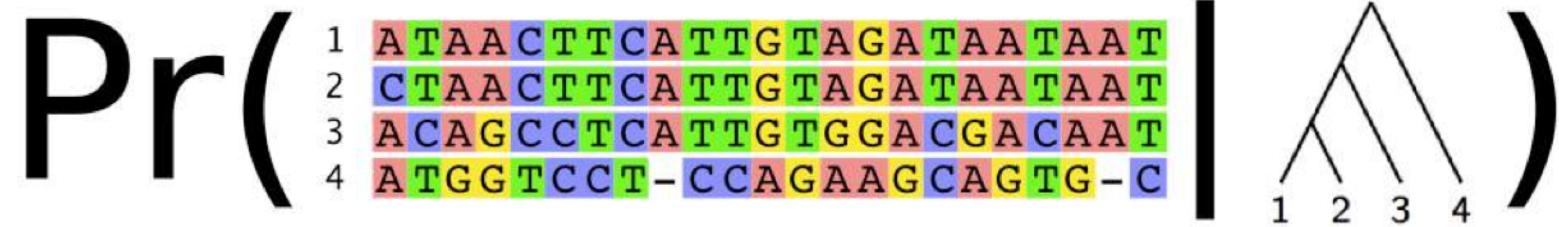
$n$	Unrooted trees ( $U_n$ )	Rooted trees ( $R_n$ )	Labelled histories ( $H_n$ )
3	1	3	3
4	3	15	18
5	15	105	180
6	105	945	2,700
7	945	10,395	56,700
8	10,395	135,135	1,587,600
9	135,135	2,027,025	57,153,600
10	2,027,025	34,459,425	2,571,912,000
20	$\sim 2.22 \times 10^{20}$	$\sim 8.20 \times 10^{21}$	$\sim 5.64 \times 10^{29}$
50	$\sim 2.84 \times 10^{74}$	$\sim 2.75 \times 10^{76}$	$\sim 3.29 \times 10^{112}$

Como podemos comparar arboles filogeneticos?

Dado un grupo de secuencias alineadas, como podemos determiner si un arbol describe/explica mejor la data que otro?

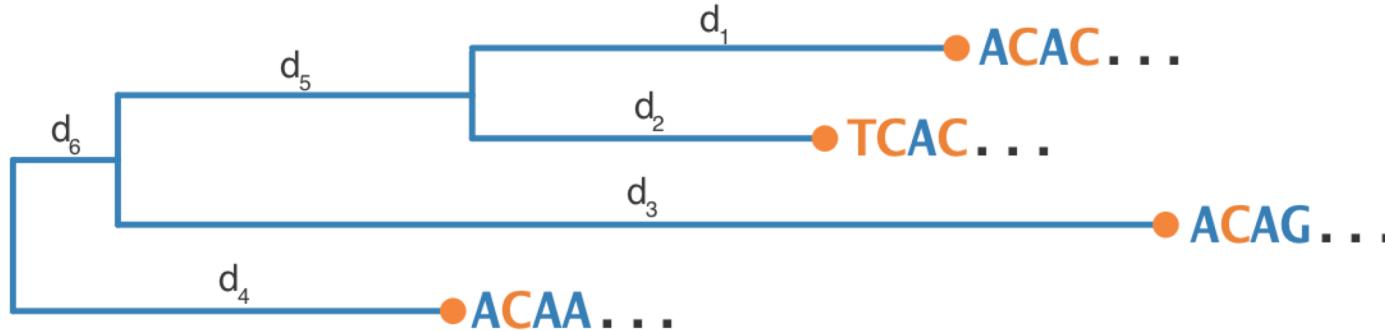
Ademas de codificar funcion, DNA sirve como un registro de historia evolutiva.

Una manera de evaluar un arbol seria calcular la probabilidad de la data dado un **modelo estadistico de evolucion de DNA**



A esta probabilidad la llamamos el **Likelihood** del arbol – el likelihood filogenetico.

Una manera de reconstruir una historia de evolutive seria encontrar un arbol que maximize el Likelihood (**Maximum Likelihood**)

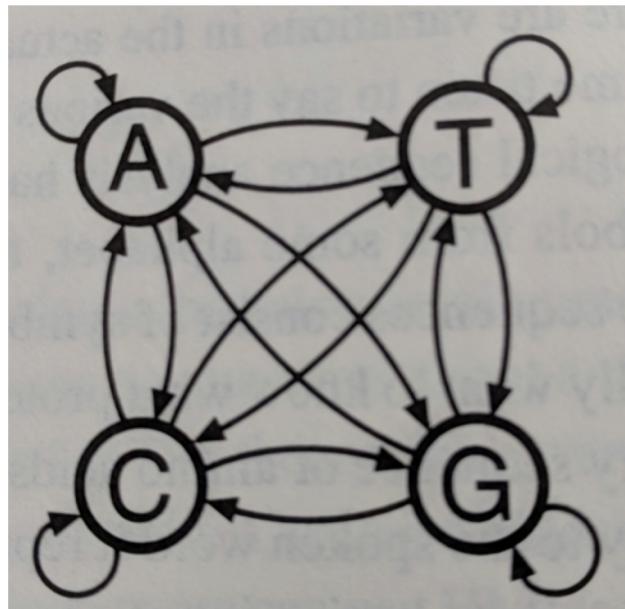


Asumimos que:

- Cada site evoluciona independientemente.
- Substituciones en cada site se rigen por un **proceso de Markov continuo**

El Modelo de substitucion:

- Vincula las sequencias genomicas con la genealogia
- Observamos las sequencias en la hojas de los arboles, no sus historias
- Substituciones multiples en el mismo site significa que no todas las substituciones son observadas.
- Para inferir la historia de evolucion necesitamos tomar en cuenta **todas las trayectorias posibles**



Expresamos el modelo de substitucion con la matriz de substitucion  $Q$  donde  $q(i,j)$  es la tasa relativa de substitucion de estado  $i$  a estado  $j$ .

$$Q = \begin{pmatrix} T & C & A & G \\ T & -(a+b+c) & a & b & c \\ C & d & -(d+e+f) & e & f \\ A & g & h & -(g+h+i) & i \\ G & j & k & l & -(j+k+l) \end{pmatrix}$$

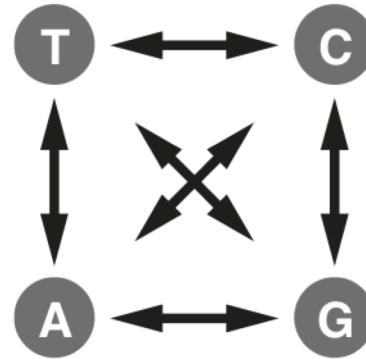
Las probabilidades de transicion en  $P(t)$  toman en consideracion cada posible traectoria evolutiva en cada site.

$$P(t) = e^{Qt}$$

$$P(t) = \begin{pmatrix} T & C & A & G \\ T & p_{tt}(t) & p_{tc}(t) & p_{ta}(t) & p_{tg}(t) \\ C & p_{ct}(t) & p_{cc}(t) & p_{ca}(t) & p_{cg}(t) \\ A & p_{at}(t) & p_{ac}(t) & p_{aa}(t) & p_{ag}(t) \\ G & p_{gt}(t) & p_{gc}(t) & p_{ga}(t) & p_{gg}(t) \end{pmatrix}$$

# Ejemplos

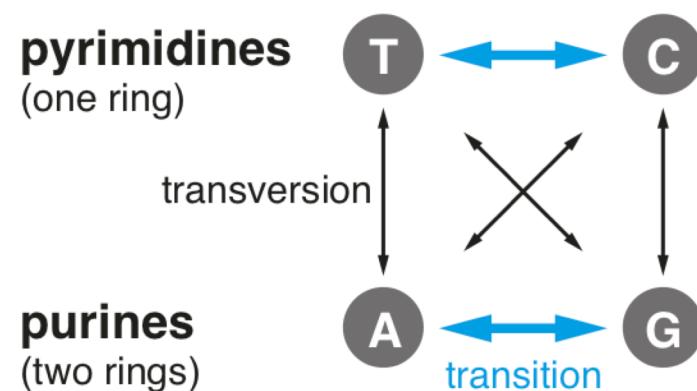
Modelo Jukes-Cantor (JC69)



$$\begin{matrix} & T & C & A & G \\ T & \cdot & \lambda & \lambda & \lambda \\ C & \lambda & \cdot & \lambda & \lambda \\ A & \lambda & \lambda & \cdot & \lambda \\ G & \lambda & \lambda & \lambda & \cdot \end{matrix}$$

$$\pi_T = \pi_C = \pi_A = \pi_G$$

**pyrimidines**  
(one ring)



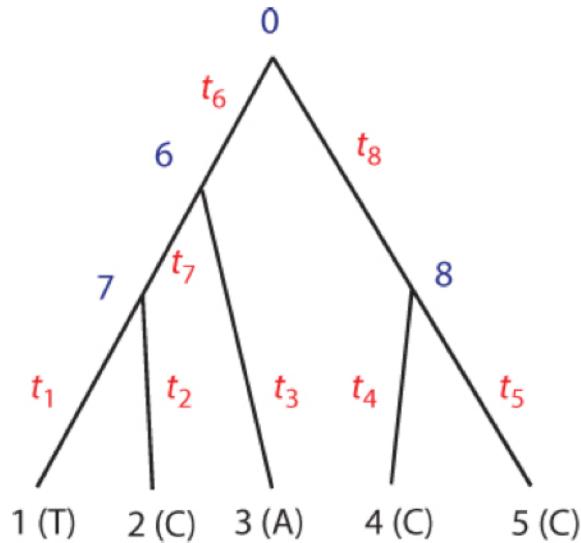
**purines**  
(two rings)

$$\begin{matrix} & T & C & A & G \\ T & \cdot & \alpha & \beta & \beta \\ C & \alpha & \cdot & \beta & \beta \\ A & \beta & \beta & \cdot & \alpha \\ G & \beta & \beta & \alpha & \cdot \end{matrix}$$

$$\pi_T = \pi_C = \pi_A = \pi_G$$

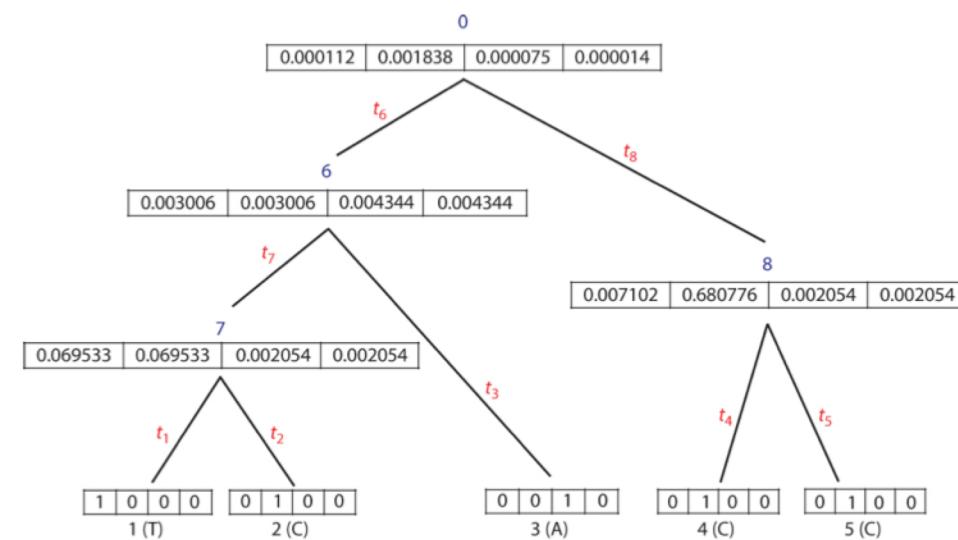
Modelo Kimura 2-parámetro (K80)

# Ejemplo Algoritmo ‘pruning’ / Felsenstein



$$f(\mathbf{x}_h \mid \theta) = \sum_{x_0} \sum_{x_6} \sum_{x_7} \sum_{x_8} [\pi_{x_0} p_{x_0 x_6}(t_6) p_{x_6 x_7}(t_7) p_{x_7 T}(t_1) p_{x_7 C}(t_2) p_{x_6 A}(t_3) p_{x_0 x_8}(t_8) p_{x_8 C}(t_4) p_{x_8 C}(t_5)].$$

$$\begin{aligned} f(\mathbf{x}_h \mid \theta) &= \sum_{x_0} \pi_{x_0} \left\{ \sum_{x_6} p_{x_0 x_6}(t_6) \left[ \left( \sum_{x_7} p_{x_6 x_7}(t_7) p_{x_7 T}(t_1) p_{x_7 C}(t_2) \right) p_{x_6 A}(t_3) \right] \right\} \\ &\times \left[ \sum_{x_8} p_{x_0 x_8}(t_8) p_{x_8 C}(t_4) p_{x_8 C}(t_5) \right]. \end{aligned}$$



Teorema de Bayes:

$$P(\theta | D) = \frac{\text{posterior likelihood} \quad \text{prior}}{\text{marginal likelihood}}$$
$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

Inferencia Bayesiana:

Metodo de inferencia estadistica que usa el teorema de Bayes para calcular la probabilidad de una hipotesis. Inferencia Bayesiana combina informacion previa con nuevas observaciones o data.

En Filogenetica queremos determinar la probabilidad de cada arbol dada informacion genetica (secuencias alineadas)

Podemos calcular la probabilidad de un arbol utilizando el Teorema de Bayes:

$$\text{Posterior probability} = \frac{\text{Likelihood} \cdot \text{Prior Probability}}{\text{Normalizing constant}}$$

Likelihood                                      Prior  
 $P(\text{Sequence} | \text{Tree}) = \Pr(\text{Sequence} | \text{Tree}) \cdot P(\text{Tree})$   
 $\Pr(\text{Sequence} | \text{Tree}) = \Pr(\text{Sequence}_1, \text{Sequence}_2, \dots, \text{Sequence}_n | \text{Tree})$   
 $= \Pr(\text{Sequence}_1 | \text{Root}) \cdot \Pr(\text{Sequence}_2 | \text{Left Child}) \cdot \Pr(\text{Sequence}_3 | \text{Middle Child}) \cdot \Pr(\text{Sequence}_4 | \text{Right Child})$

Utilizando MCMC (algoritmo Markov chain Monte Carlo) Podemos producir un sampleo de arboles de esta distribucion a-posteriori (posterior probability) sin necesidad de calcular la likelihood marginal (constante normalizadora en el denominador)

## BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis

Remco Bouckaert , Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, Alexei J. Drummond  [ view less ]

Autores provienen de 18 instituciones provenientes de 9 países

BEAST 2.5 permite construir una gran variedad de modelos filogeneticos para ser aplicados a secuencias alineadas y data comparativa (utilizando la interface grafica BEAUti).

BEAST estima/trabaja con **time-trees** / arboles de tiempo.

Los modelos filogeneticos estan compuestos por:

- La distribucion a-priori de time-trees (coalescente o modelos birth-death)
- El modelo de substitucion (matriz de substitucion)
- El site-model (como el modelo de substitucion varia entre locis/loci)
- El modelo de reloj molecular (estricto, relajado, random local clock)
- Como asignar las distintas partes del modelo si la data contiene particiones.

Distintos modelos son manejados/diferenciados con top-level templates (BEAUti).

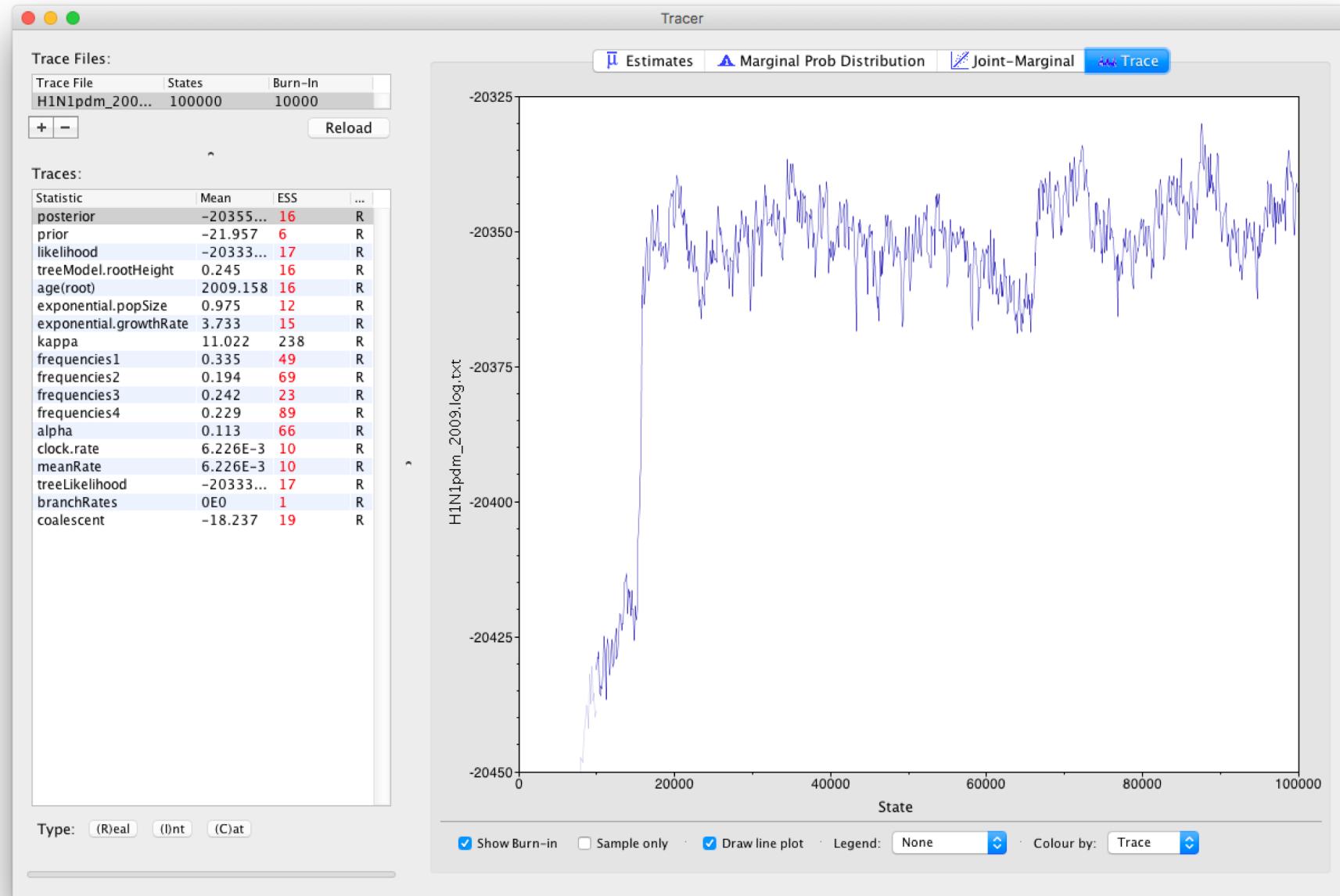
Nuevos sub-modelos y templates pueden ser disenados por terceros (desarrolladores de software). BEAST 2.5 es extensible.

BEAST produce archivos de registro (log files) reportando los valores 'sampleados' durante la ejecucion del MCMC.

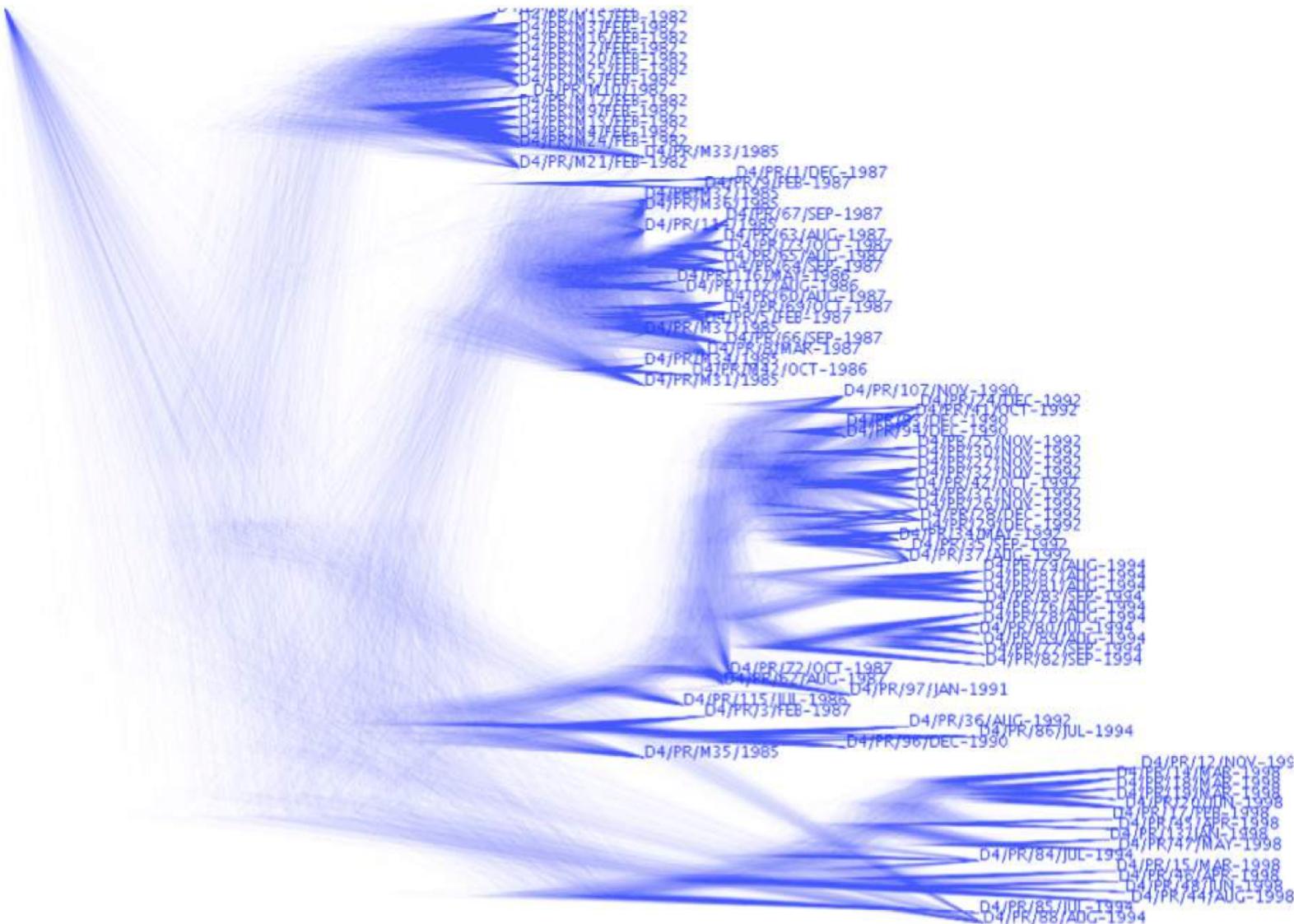
No todas las muestras son reportadas: primero se eliminan las muestras generadas durante la etapa de burn-in. El usuario ademas especifica la frecuencia con la que se deben de reporter las muestrar para evitar correlacion entre ellas.

Los valores en los logs representan las distribuciones a-posteriori de los arboles y todos los parametros e hiper-parametros utilizados en el analisis. Ademas se pueden reporter otros valores como Altura de la raiz, ancestral state probabilities, etc.

Con ellos podemos generar arboles de consenso y estadisticas. La data generada puede ser utilizada para otros analisis.



# BEAST 2.5 Distribucion a-posteriori de arboles



# Ejemplo: Fiebre Amarilla en las Americas



- La hipótesis más citada sobre el origen de la fiebre amarilla en las Américas especula que el virus fue introducido de África, junto con los mosquitos *Aedes aegypti*, en los barcos utilizados durante el tráfico de esclavos.
- Esta hipótesis no había sido rigurosamente examinada, previo a Bryant et al. (2017), utilizando información genética (secuencias) y técnicas filogenéticas modernas para la estimación de tiempos/fechas de divergencia.
- Bryant et al proveen la primera evidencia directa que la YFV fue introducida durante la época del tráfico de esclavos.

## Out of Africa: A Molecular Perspective on the Introduction of Yellow Fever Virus into the Americas

Juliet E Bryant , Edward C Holmes, Alan D. T Barrett 

Published: May 18, 2007 • <https://doi.org/10.1371/journal.ppat.0030075>

Data set: 133 YFV prM/E gene sequences (human, mosquito, monkey) from 22 countries (14 African, 8 South American)

BEAST analysis:

- chain length 25 million, sampling every 1000.
- GTR+I+Gamma4 / rate variation among lineages / relaxed clock

The phylogenetic analysis infers a time-scale and evolutionary history of YFV.

It estimates:

- The rate of molecular evolution
- The date of the most recent common ancestor

# Fiebre Amarilla (YFV) en las Americas

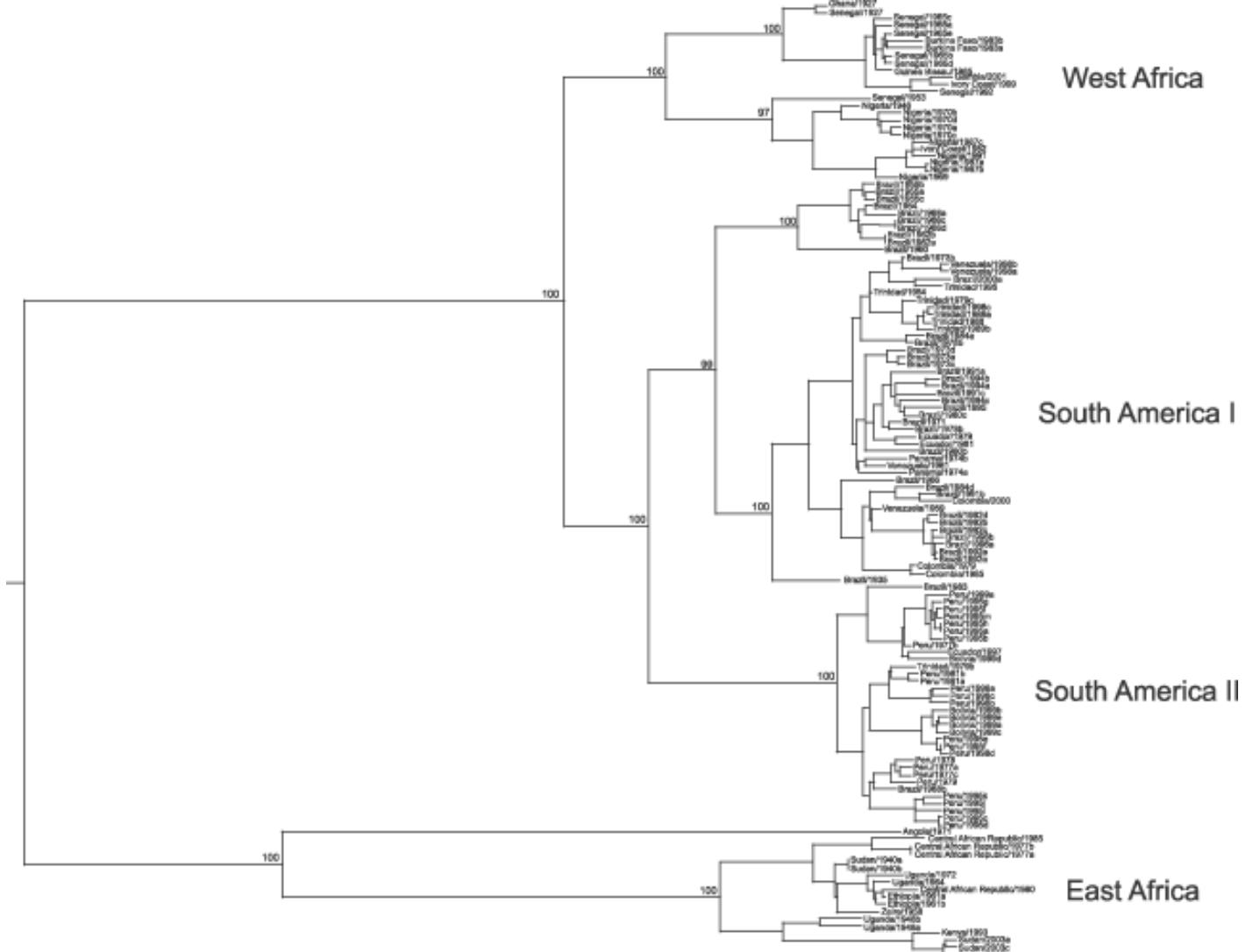
YFV puede ser dividida en dos grupos geográficos, con distintas linajes virales observados en África y las Américas

Los casos (isolates) Americanos son monofileticos.

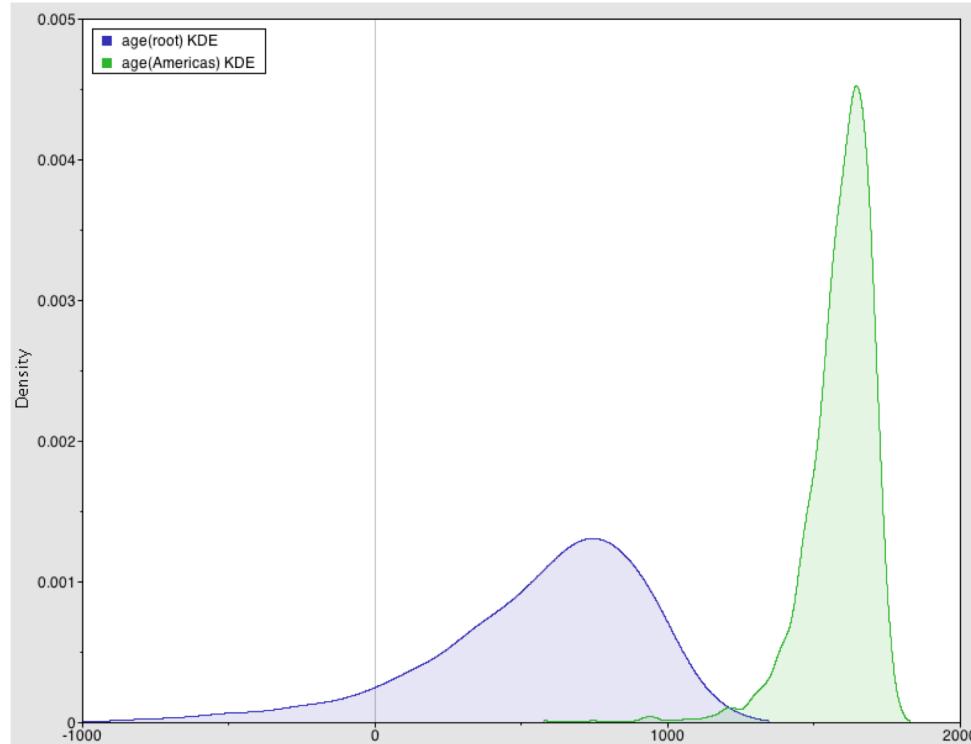
Los casos Americanos se dividen en aquellos que provienen del este y oeste del continente.

Los casos de Africa Occidental estan mas cercanamente relacionados que aquellos de las Americas.

Los casos de África del Este son más divergentes.



A Posteriori de las fechas de la raiz and los clades Sud Americanos\*



## Estimated divergence times (years)

Edad Raiz (origen de todas las muestras):  
media 723 [288,1304] 95% HPD

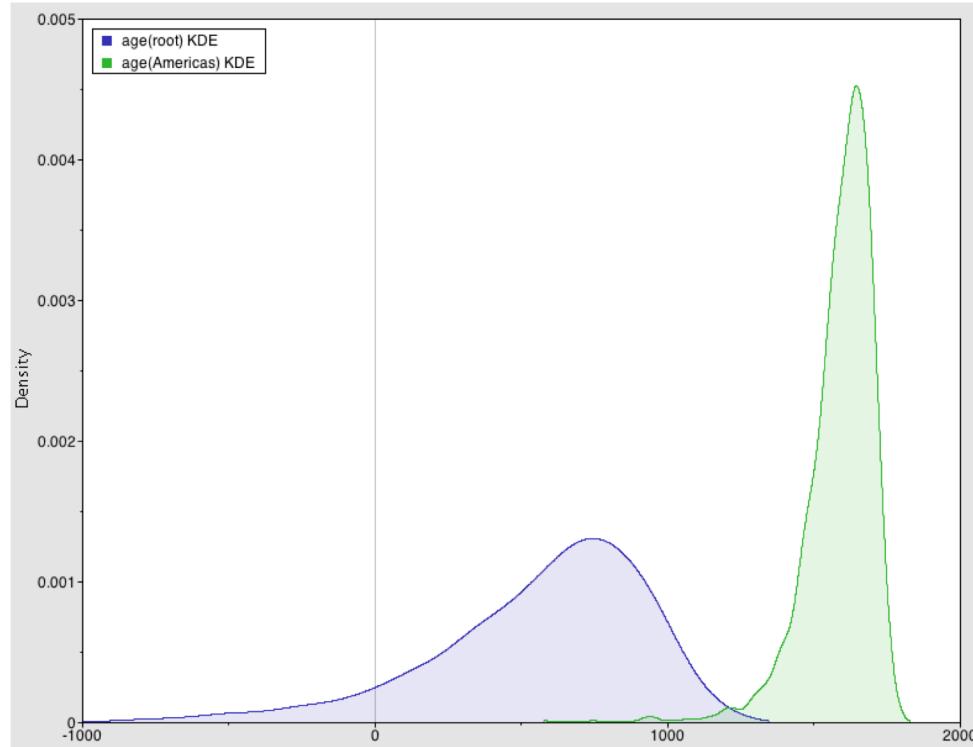
Divergencia S.America and Africa Occidental:  
media 470 [186,869] 95% HPD

Origen de los genotipos Sud Americanos:  
media 306 [120,590] 95% HPD

Muestra mas reciente = 2007

\*Extracted from the *Estimating rates and dates from time-stamped sequences BEAST tutorial*. It uses a subset of the original sampled sequences. 2009 time of most recent sample

A Posteriori de las fechas de la raiz and los clades Sud Americanos\*



## Estimated divergence times (years)

Edad Raiz (origen de todas las muestras):  
media 723 [288,1304] 95% HPD

Divergencia S.America and Africa Occidental:  
media 470 [186,869] 95% HPD

Origen de los genotipos Sud Americanos:  
media 306 [120,590] 95% HPD

Muestra mas reciente = 2007

- La diversidad genética de las muestras disponibles de YFV en América del Sur surgió hace 3 a 4 siglos.
- Existe una fuerte evidencia que apoya la teoría de una introducción inicial durante el período de tráfico de esclavos y el primer contacto entre los dos continentes.

## The early spread and epidemic ignition of HIV-1 in human populations

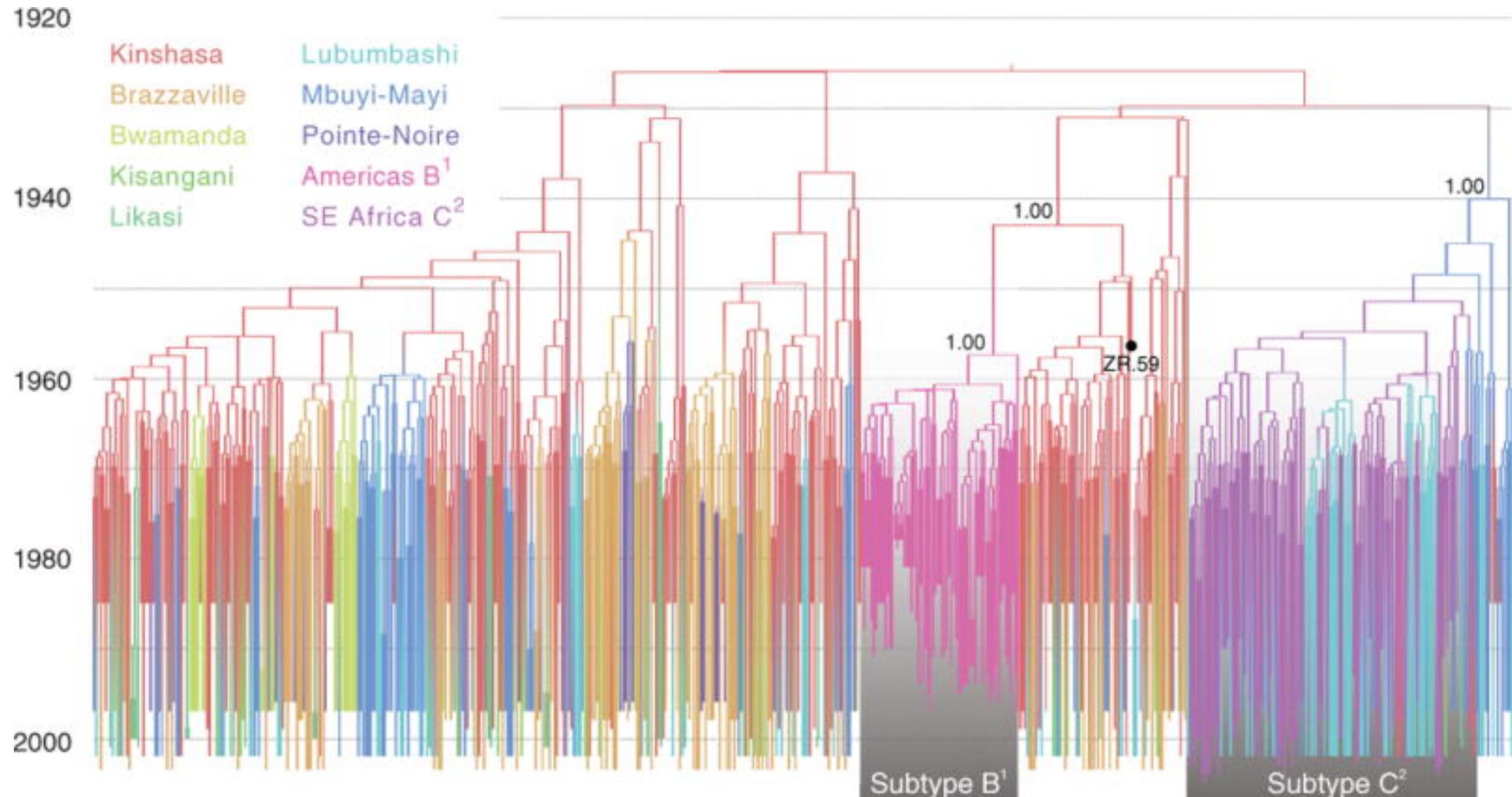
Nuno R. Faria,<sup>1,2</sup> Andrew Rambaut,<sup>3,4,5</sup> Marc A. Suchard,<sup>6,7</sup> Guy Baele,<sup>2</sup> Trevor Bedford,<sup>8</sup> Melissa J. Ward,<sup>3</sup>  
Andrew J. Tatem,<sup>4,9</sup> João D. Sousa,<sup>2,10</sup> Nimalan Arinaminpathy,<sup>1</sup> Jacques Pépin,<sup>11</sup> David Posada,<sup>12</sup>  
Martine Peeters,<sup>13</sup> Oliver G. Pybus,<sup>1,\*†</sup> and Philippe Lemey<sup>2,\*†</sup>

# Orígenes del VIH

MRC

Centre for  
Global Infectious  
Disease Analysis

Imperial College  
London

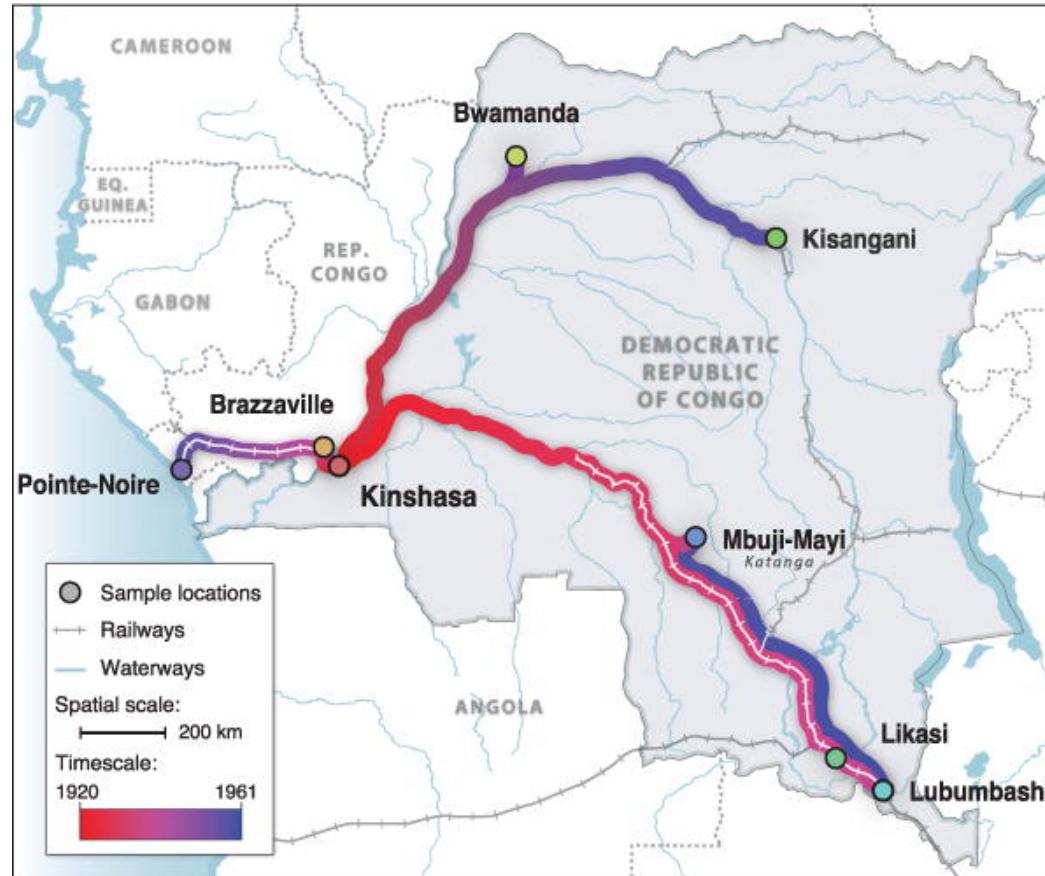


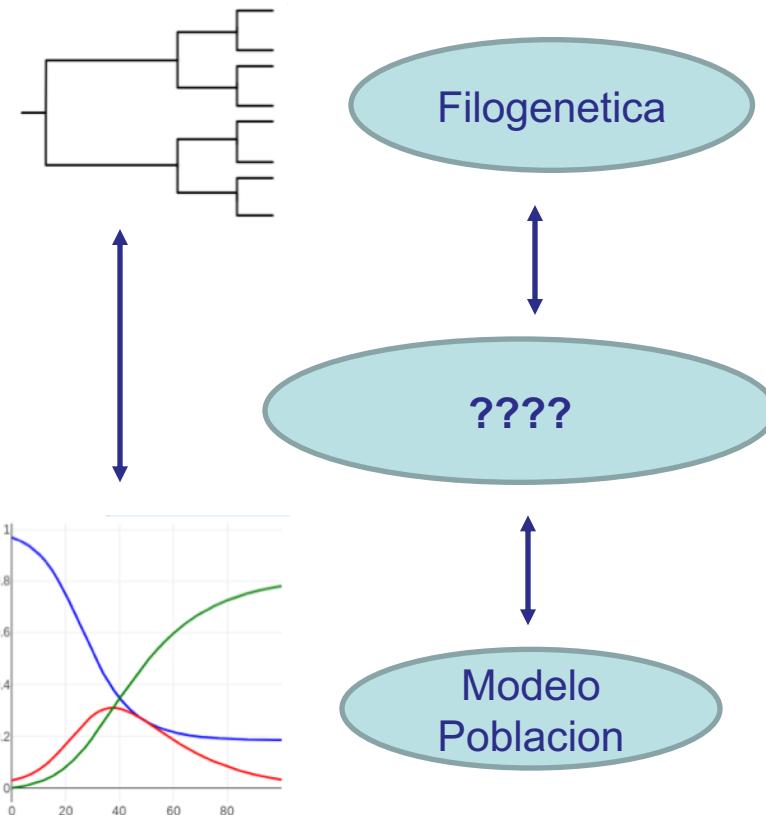
# Orígenes del VIH

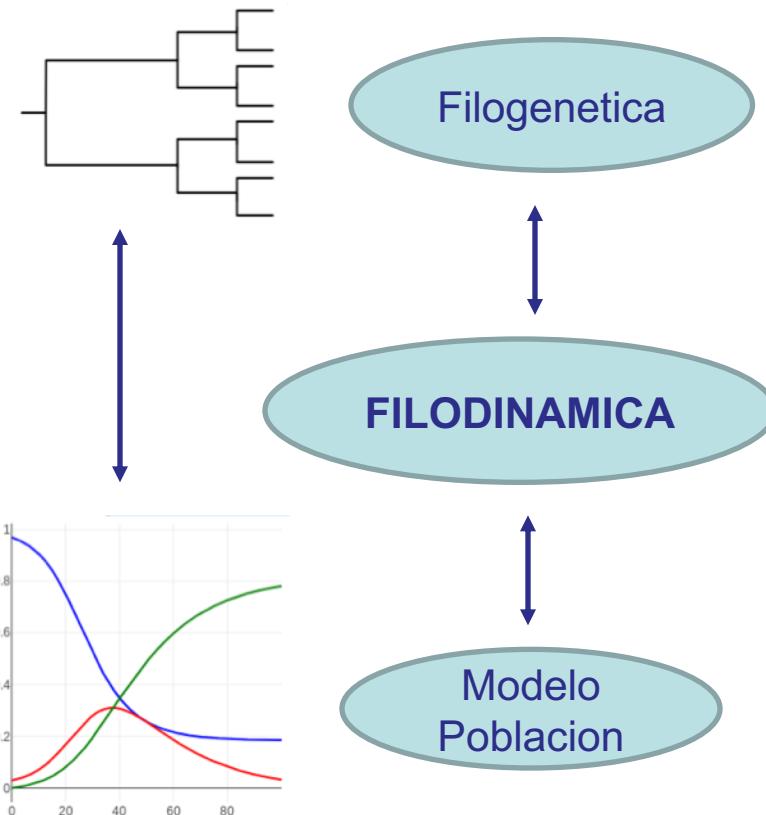
MRC

Centre for  
Global Infectious  
Disease Analysis

Imperial College  
London







**La Filodinámica** vincula propiedades filogenéticas de patógenos con dinamicas poblacionales eg. historia / trajecotria de epidemia / brote.

Permite la estimacion de parametros epidemiologicos desde una muestra de secuencias de patogenos

La inferencia filodinamica require relacionar un modelo del proceso epidemiologico con un proceso genetico poblacional.

Procesos: simulacion directa, modelos.coalescentes, modelos sampling-birth-death

BEAST ofrece métodos **filodinámicos** bajo la forma de distribuciones a-priori de árboles (Tree Priors). Estas distribuciones asignan un valor de probabilidad a cada árbol posible. Las podemos separar en dos tipos:

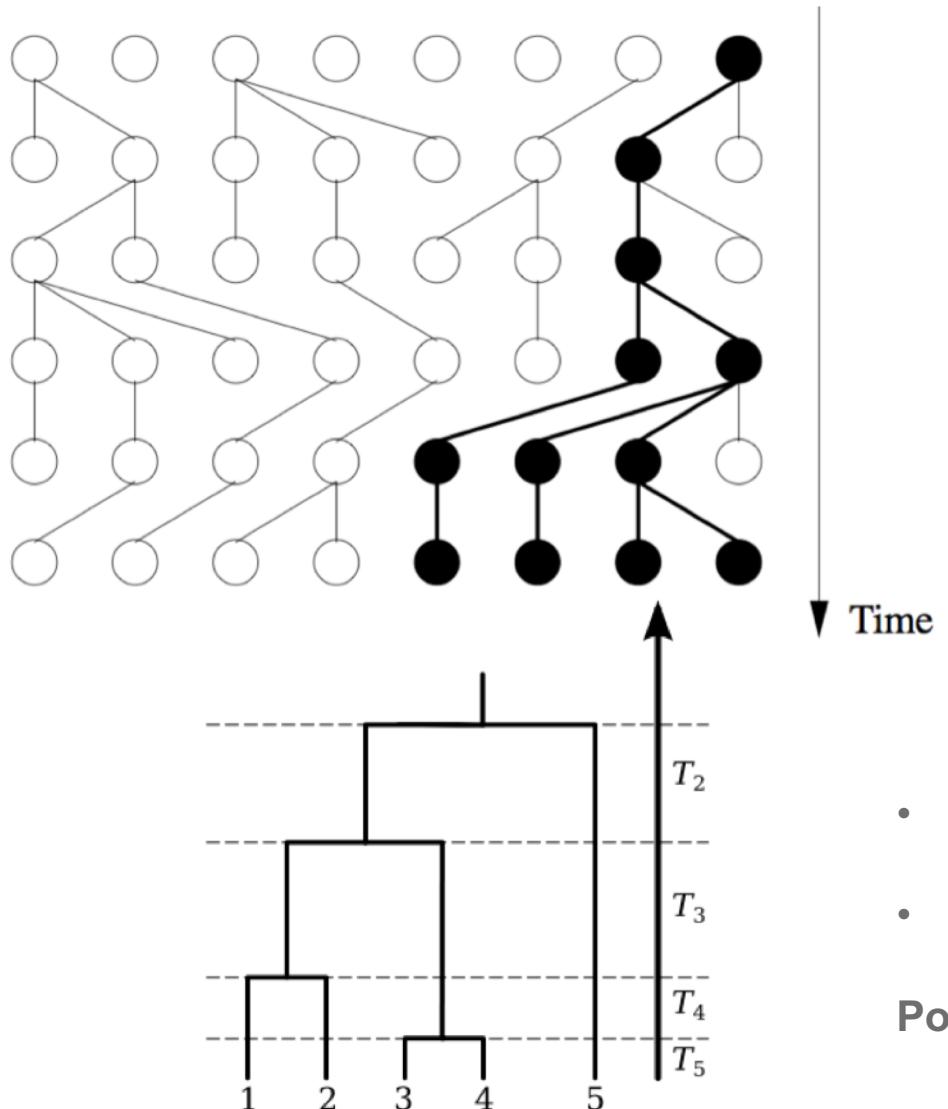
Procesos de Nacimiento-Muerte (Birth-Death processes):

- Modelo Yule: proceso de Nacimiento puro / tasa de nacimiento constante  $\lambda$ .
- Extensiones: tasas constantes de Nacimiento-Muerte / fracción de muestreo.

Proceso de Coalescencia / Coalescent

- No/Semi-paramétricos
  - Bayesian Skyline Plot (BSP) y variaciones.
- Parámetricos
  - Población Constante / Modelo exponencial
  - Structured Coalescent

- El Coalescente es un modelo estocastico que describe la (generacion de la) genealogia / ancestry de una muestra de una poblacion.
- El Coalescente esta basado en el idealizado modelo poblacional Wright-Fisher:
  - poblacion constante,
  - generaciones discretas,
  - las generaciones no se superponen,
  - La generacion  $t+1$  esta formada de la generacion  $t$  con sampleo uniforme con reemplazo
- El modelo Wright-Fisher

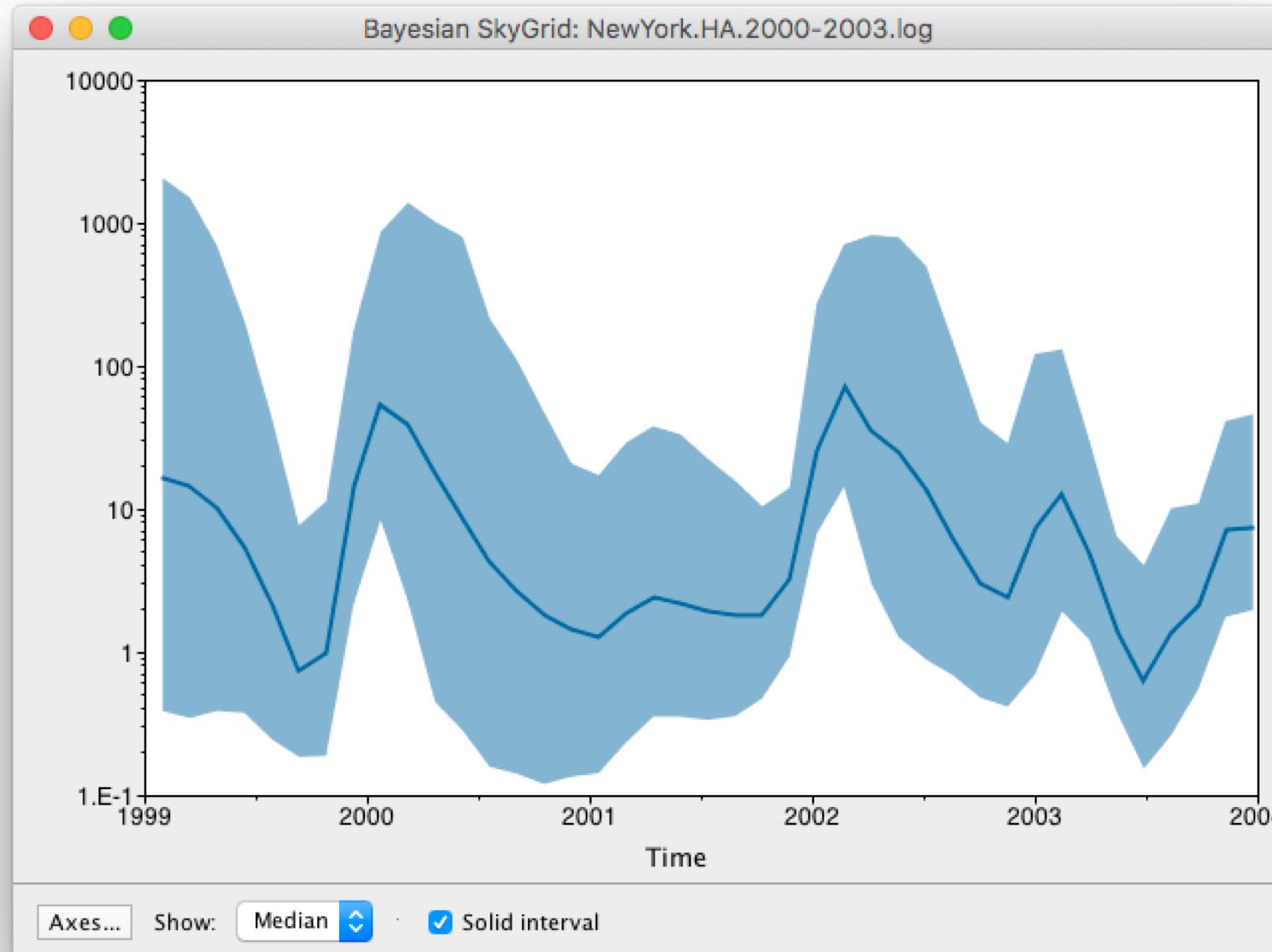


- El Coalescente es un modelo estocastico que describe la (generacion de la) genealogia / ancestry de una muestra de una poblacion.
- El Coalescente esta basado en el idealizado modelo poblacional Wright-Fisher:
  - poblacion constant ( $N$ ), generaciones discretas
  - las generaciones no se superponen,
  - La generacion  $t+1$  esta formada de la generacion  $t$  con sampleo uniforme con reemplazo
- El modelo Wright-Fisher, para valores grandes de  $N$ , genera una distribucion de time-trees: la distribucion de coalescent time-trees.

- Tasa par de coalescencia:  $\lambda(t) = 1/N_e(t)$
- Tasa de coalescencia total:  $\binom{A(t)}{2} / N_e(t)$

**Podemos calcular el Likelihood!!**

# Skyline Bayesiano (reconstrucción)

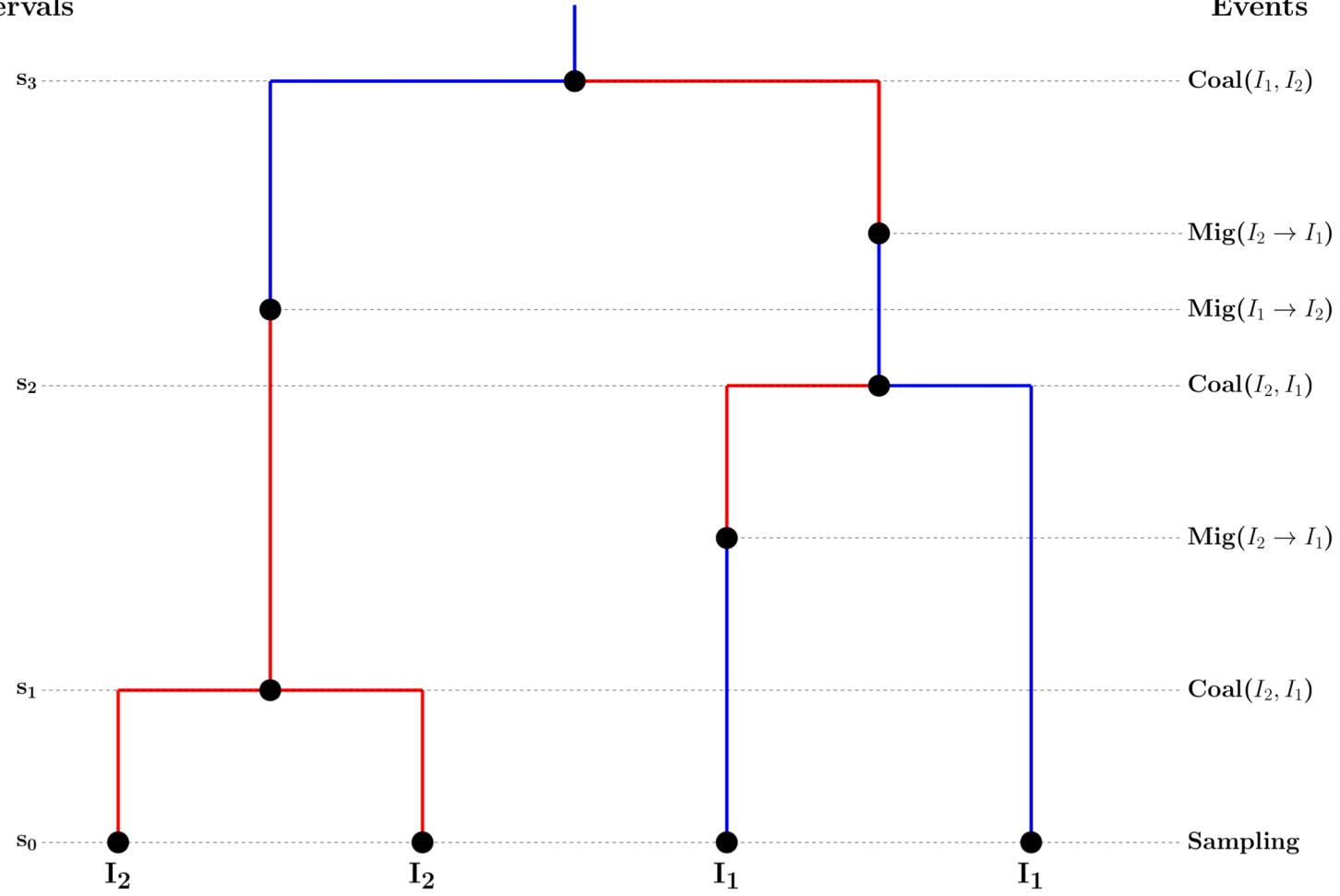


## Priors no parametricos Bayesian Skyline methods

Asumir cambios de poblaciones en partes, entre eventos de coalescencia i.e. intervalos en el arbol.

Los cambios de poblacion se pueden modelar de distintas maneras: SkyGrid, SkyRide

## Intervals



## Events

Coal( $I_1, I_2$ )

Mig( $I_2 \rightarrow I_1$ )

Mig( $I_1 \rightarrow I_2$ )

Coal( $I_2, I_1$ )

Mig( $I_2 \rightarrow I_1$ )

Coal( $I_2, I_1$ )

Sampling

El coalescente es un modelo estocastico que describe la (generacion de la) genealogia / ancestry de una muestra de una poblacion.

El coalescente estructurado trabaja con poblaciones estructuradas.

Es util imaginar a los procesos epidemiologicos como **birth-death-migration processes**:

- Eventos de coalescencia corresponden a nacimientos / infeccion.
- Muertes corresponden a mortalidad o recuperacion.
- Eventos de migracion corresponden a individuos en transicion entre sub-poblaciones / compartimientos en modelos tipo SIR.

# Bayesian phylodynamic inference with complex models

Erik M. Volz , Igor Siveroni

PLOS Computational  
Biology  
Nov 2018

PhDyn es un paquete de BEAST2.5 para realizar inferencia bayesiana filogenetica utlizando modelos con **poblacion estructurada y dinamica compleja**. PhyDyn permita estimar simultaneamente arboles filogeneticos y parametros epidemiologicos utilizando secuencias moleculares.

PhDyn implementa el **coalescente estructurado** (Volz, 2012) para una gran variedad de procesos poblacionales (epidemias).

Input:

- Modelos poblacionales (epidemias) son especificados con ecuaciones diferenciales (**ODEs**) describiendo la dinamica poblacional de demes y no-demes. Esta dinamica es modelada utilizando una combinacion de nacimientos (birth matrix F i.e. infecciones), migraciones (migration matrix G) y muertes (death vector  $\mu$ ).
- Arboles filogeneticos anotados con informacion indicando la **subpoblacion** a la que pertenecen las muestras (tips).

PhDyn calcula:

- Trayectoria de la poblacion i.e solucion de ODEs: *timeseries* con tamaños de poblacion y matrices de tasas (F,G)
- Las tasas de coalescencia y *state probabilities* en cada intervalo del arbol: probabilidad que un punto del linaje, en un tiempo t, ocupe un estado (deme) determinado.
- El log-likelihood del arbol dada la trayectoria/historia poblacional.

$$\dot{I}_0 = S(\beta_0 I_0 + \beta_1 I_1) - \gamma_0 I_0$$

$$\dot{I}_1 = \gamma_0 I_0 - \gamma_1 I_1$$

$$\dot{S} = bS - S(\beta_0 I_0 + \beta_1 I_1)$$



$$F(t) = \begin{pmatrix} \beta_0 I_0(t)S & 0 \\ \beta_1 I_1(t)S & 0 \end{pmatrix} \quad G(t) = \begin{pmatrix} 0 & \gamma_0 I_0(t) \\ 0 & 0 \end{pmatrix} \quad \mu(t) = \begin{pmatrix} 0 \\ \gamma_1 I_1(t) \end{pmatrix}$$

$$\dot{E} = \beta(t)I_l(t) + \tau\beta(t)I_h(t) - (1 - p_h)\gamma_0 E(t) - p_h\gamma_0 E(t)$$

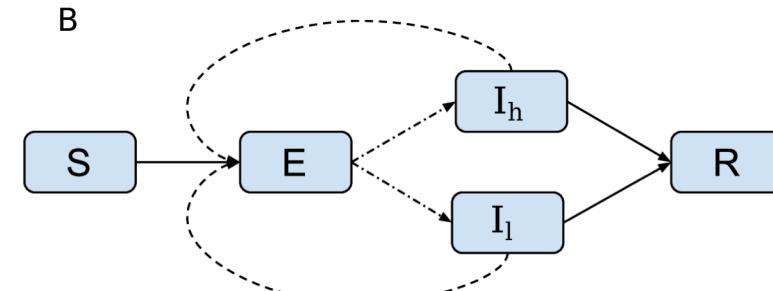
$$\dot{I}_l = (1 - p_h)\gamma_0 E(t) - \gamma_1 I_l(t)$$

$$\dot{I}_h = p_h\gamma_0 E(t) - \gamma_1 I_h(t)$$



$$\text{demes} = \{E, I_l, I_h\} \quad \beta(t) = at + b \quad \mu(t)^T = (0 \ I_l(t) \ I_h(t))$$

$$F(t) = \begin{pmatrix} 0 & 0 & 0 \\ \beta(t)I_l(t) & 0 & 0 \\ \tau\beta(t)I_h(t) & 0 & 0 \end{pmatrix} \quad G(t) = \begin{pmatrix} 0 & (1 - p_h)\gamma_0 E(t) & p_h\gamma_0 E(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$



Intervals

$s_3$

$s_2$

$s_1$

$s_0$

$I_2$

$I_2$

$I_1$

$I_1$

Events

Coal( $I_1, I_2$ )

Mig( $I_2 \rightarrow I_1$ )

Mig( $I_1 \rightarrow I_2$ )

Coal( $I_2, I_1$ )

Mig( $I_2 \rightarrow I_1$ )

Coal( $I_2, I_1$ )

Sampling

$t$

$S(t) \ I1(t) \ I2(t)$

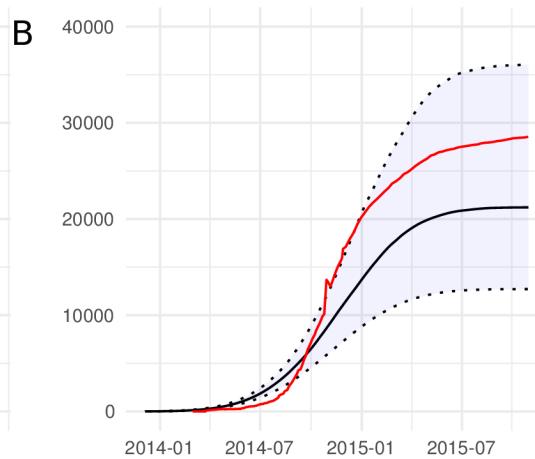
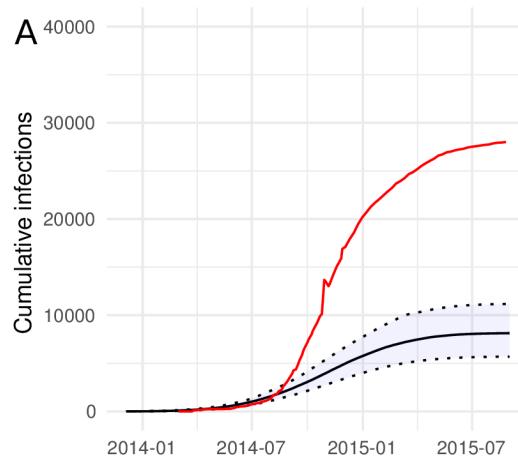
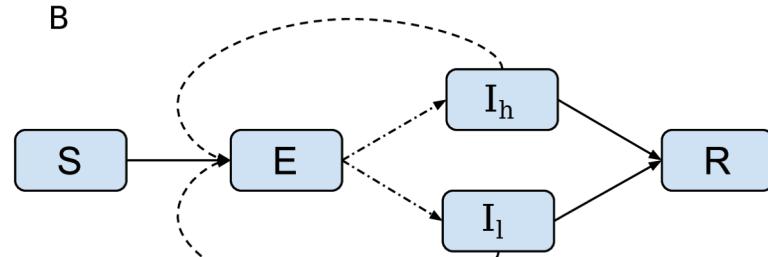
$F(t) \ G(t) \ D(t)$

$[ p_i(I1), p_i(I2) ]$

$\lambda_{i,j}(t) \ A(t)$

Likelihood

Solucion ODE: Trayectoria



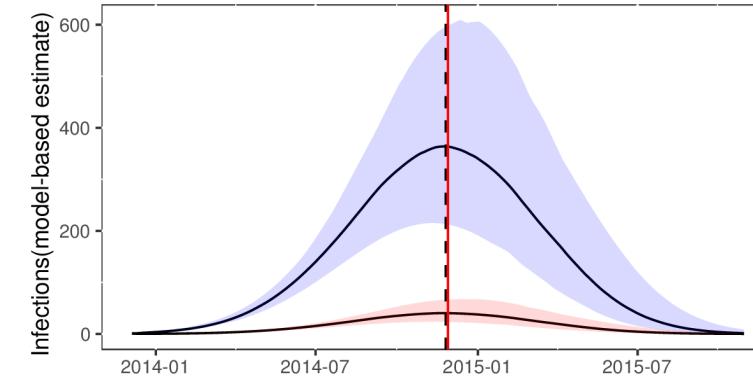
Numero de infecciones acumuladas estimadas  
utilizando un SEIR simple y SEI2R. Read line plots  
WHO reported cases

## Superspreading SEIR: SEI2R

Modelo SEIR con dos demes.

Deme  $I_h$  de alto riesgo.

Arbol de maxima credibilidad  
utilizando arboles generados por  
Dudas et.al

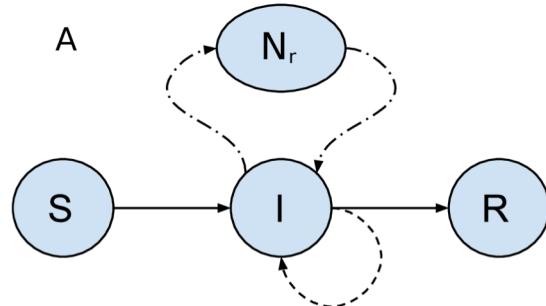


$$R_0 = 1.52 [1.48 - 1.54]$$

Estimated peak: Nov. 25 2014

WHO peak: Nov 28 2014

10% genera 43-54% de nuevos casos



**Caso:** Virus Influenza A (IAV) H3N2.

Estacion 2004-05.

**Data:** 102 HA-1 sequencias recolectadas entre 2004 y 2005 en el estado de NY.

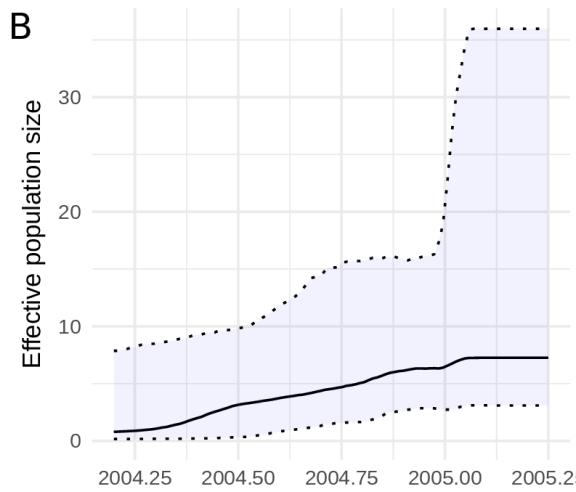
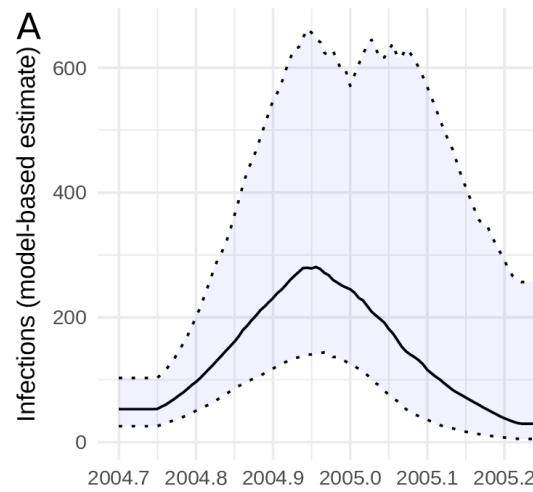
**Modelo:** SIR simple + reservorio global de IAV.

Demes: I / Nr

**R0**

Tasa de Reproduccion Basica estimada  
 $R_0 = 1.16 [1.07 - 1.30]$

$R_0 = 1.22$  para 2004-05 NY epidemic  
(Bettancourt et.al  $R_0=1.22$ )



**(A)** Numero de infecciones estimadas usando PhyDyn  
**(B)** Bayesian Skyline Plot convencional.

La fecha del pico de incidencia es correctamente identificado: final del 2004

El BSP no detecta ni el pico ni la disminucion de prevalencia al final de la estacion.

BEAST1: <http://beast.community>

BEAST2: <https://www.beast2.org>

Taming the BEAST: <https://taming-the-beast.org>

PhyDyn: <https://github.com/mrc-ide/PhyDyn>

PhyDyn wiki: <https://github.com/mrc-ide/PhyDyn/wiki>

Gracias