

# Filogenética molecular

Bioinformática para biotecnología BIT120

31 marzo 2017

Eduardo Castro, PhD

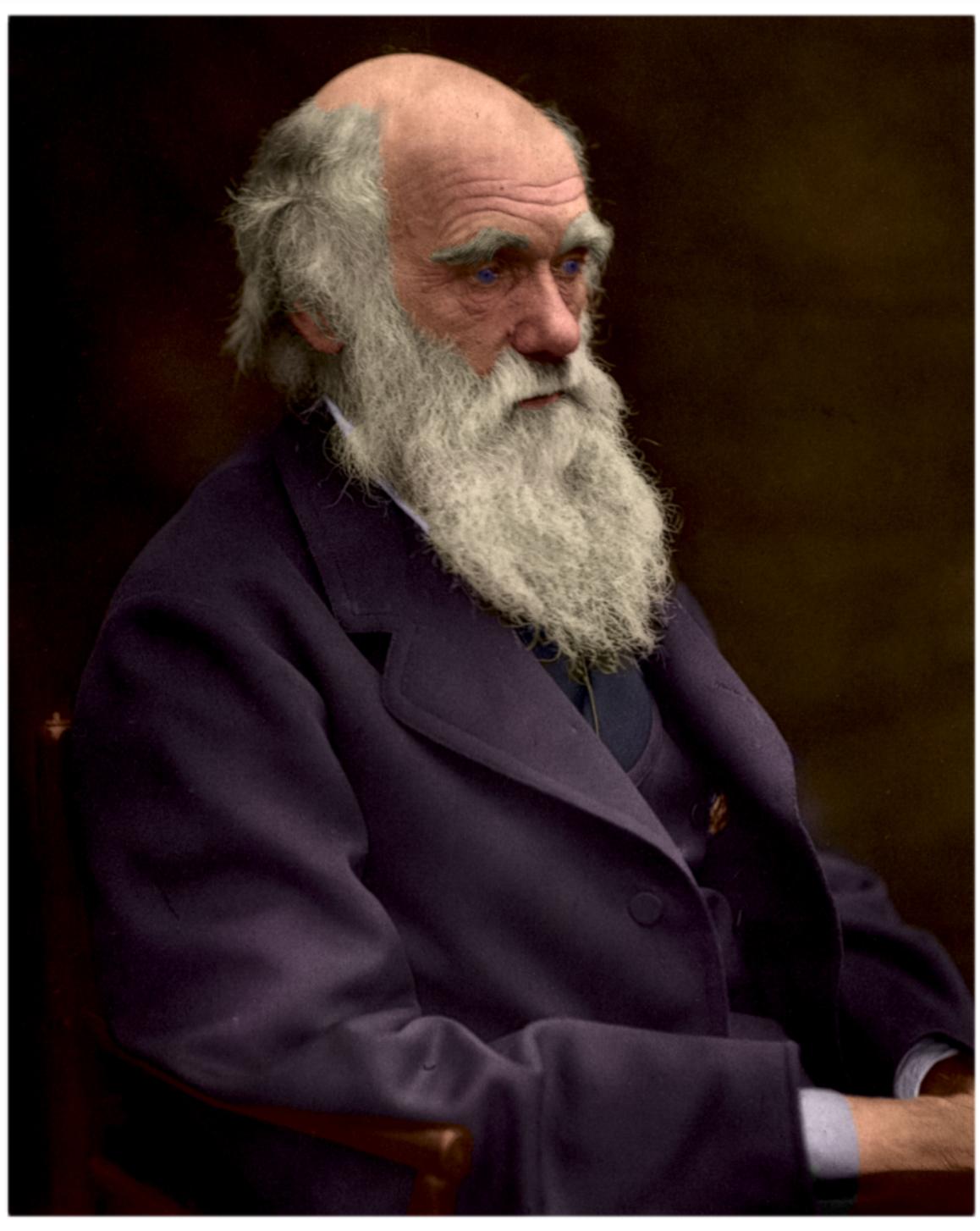
Center for Bioinformatics and Integrative Biology

[www.castrolab.org](http://www.castrolab.org)

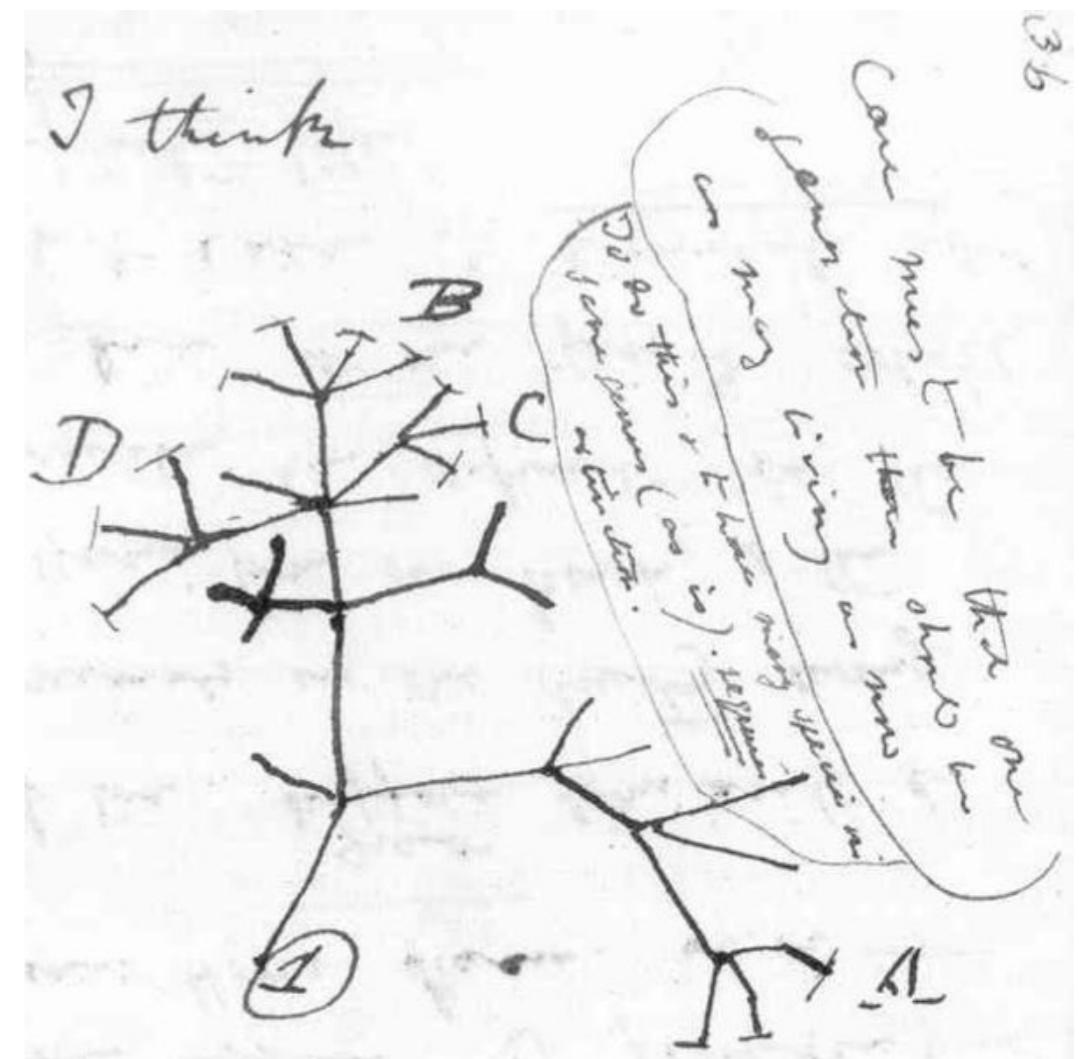
# ¿Qué es una filogenia?

- Es una hipótesis sobre la evolución de un grupo de secuencias
- Recordar alineamientos múltiples —> hipótesis de homología
- Puede ser utilizado para entender la evolución de especies, e.g., tree of life
- Sinónimos = filogenia, árbol filogenético, árbol

# ¿Qué es una filogenia?



...from so simple a beginning endless forms most beautiful  
and most wonderful have been, and are being, evolved  
The origin of species, 1859



de un cuaderno de Darwin de 1837

# ¿Qué es una filogenia?

- Una explicación de cómo las secuencias han evolucionado, sus relaciones genealógicas, y por lo tanto de cómo han llegado a ser de la forma que son hoy día

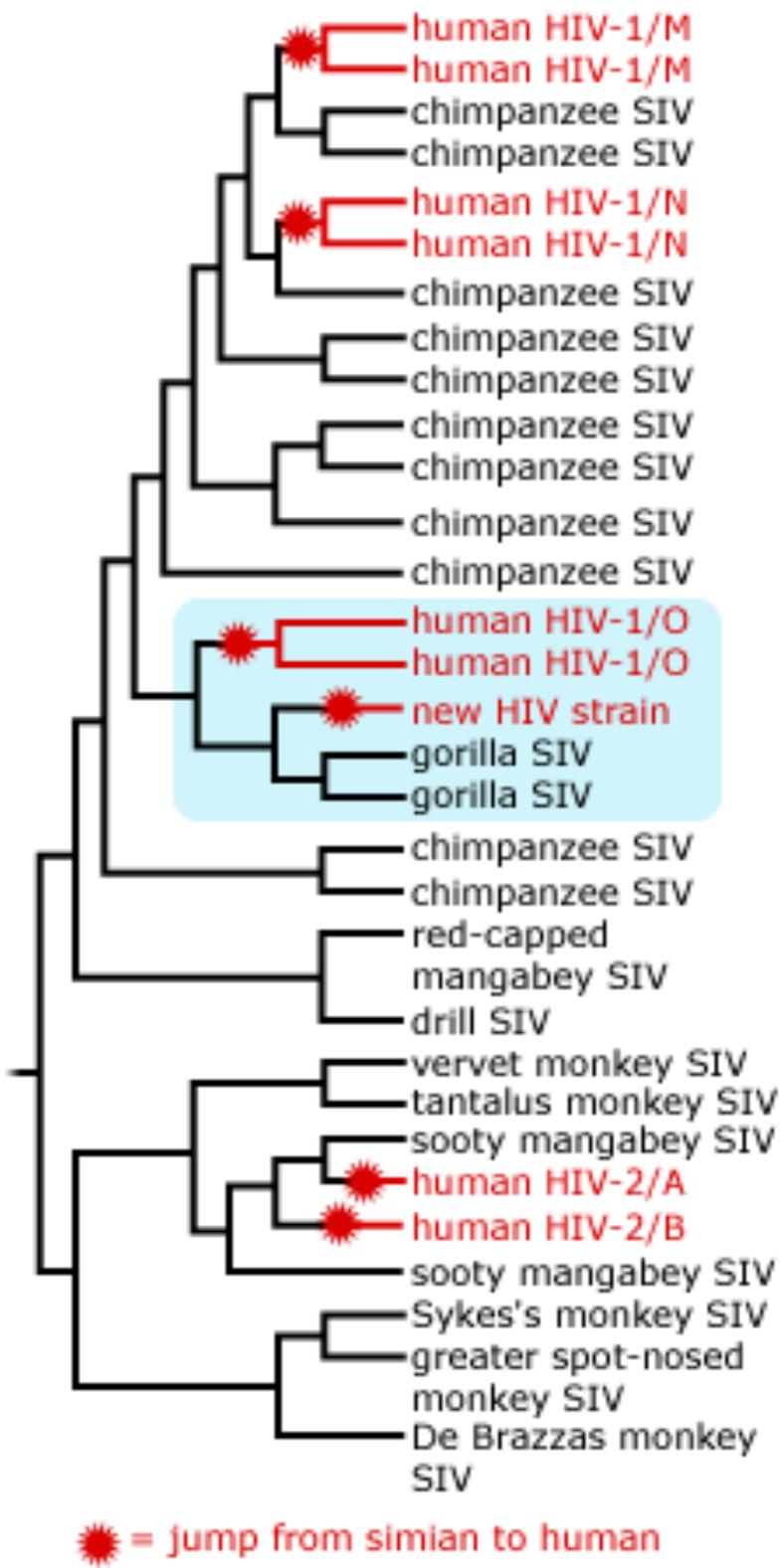
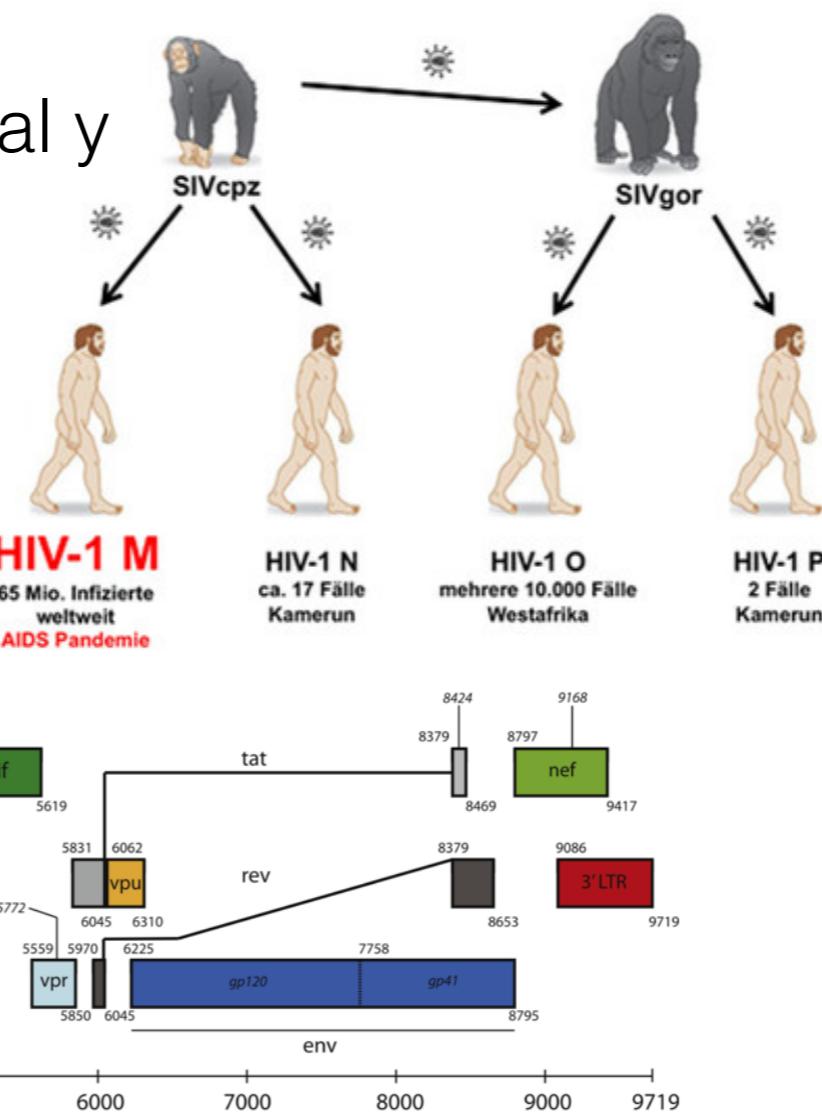
\*Pregunta de prueba

# ¿Para qué sirven?



# Ejemplo sobre la utilidad de las filogenias

- Entender zoonosis
- Entender bajo qué condiciones ocurren
- Seguir epidemias temporal y espacialmente

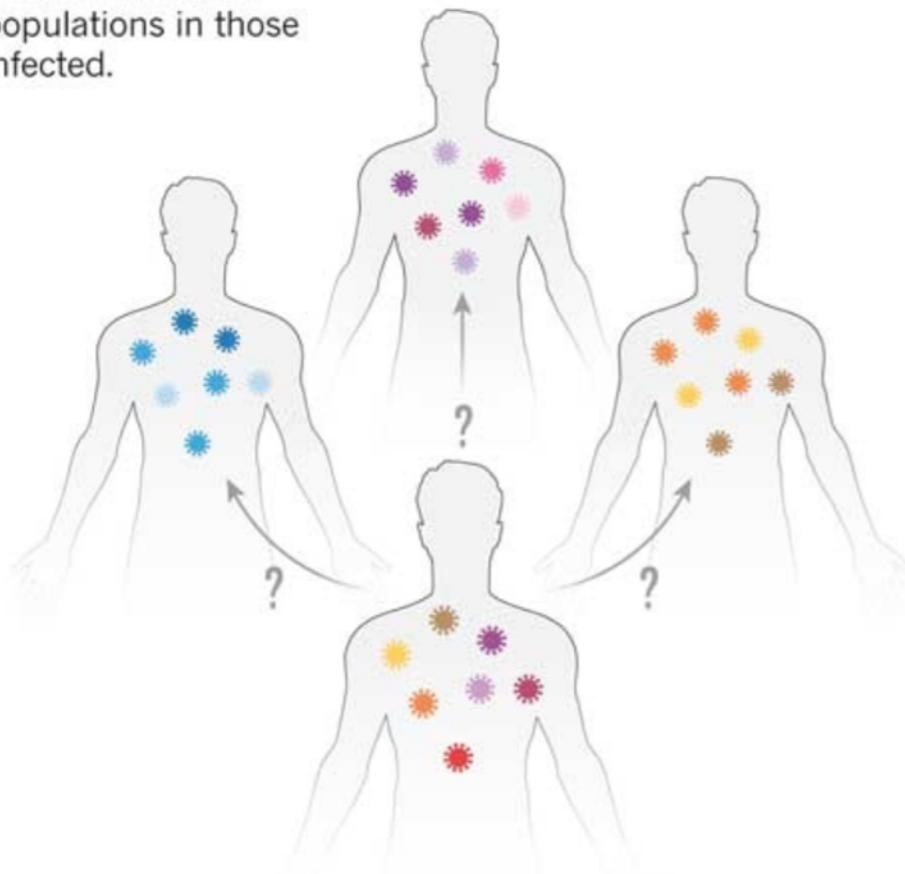


# Filogenias como evidencia forense

## Infectious forensics

Phylogenetics offers a way to establish relationships between microbes infecting several individuals and can be used as corroborating evidence when someone is suspected of infecting others with a disease.

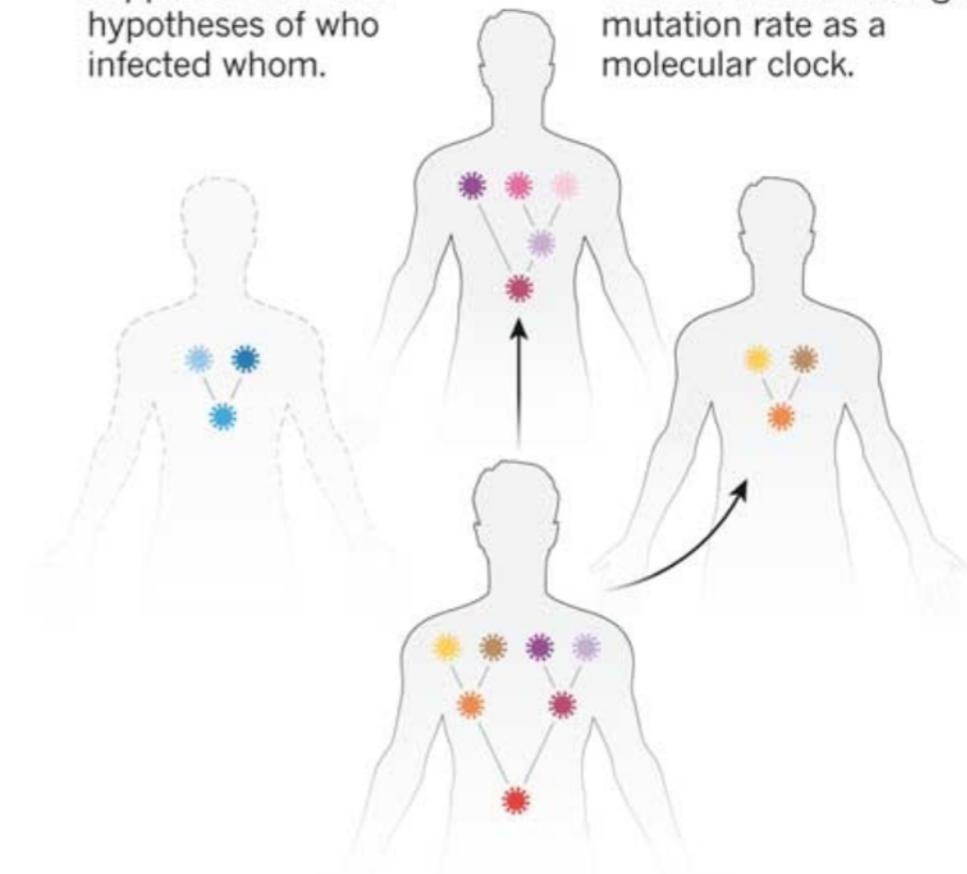
- 1 Pathogen genomes can mutate quickly, creating diverse microbial populations in those infected.



- 2 By sequencing highly variable regions of pathogen genomes scientists can build a phylogenetic tree that suggests how the microbes are related.



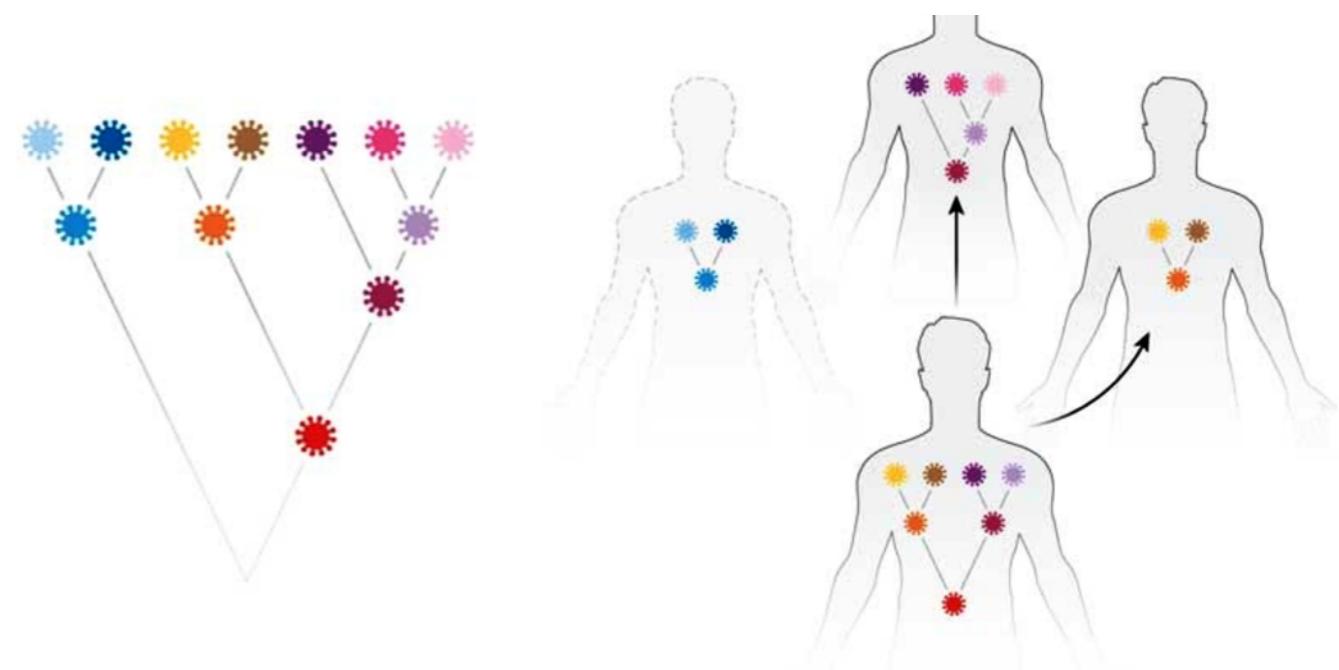
- 3 The relatedness of the viral populations can support or rule out hypotheses of who infected whom.



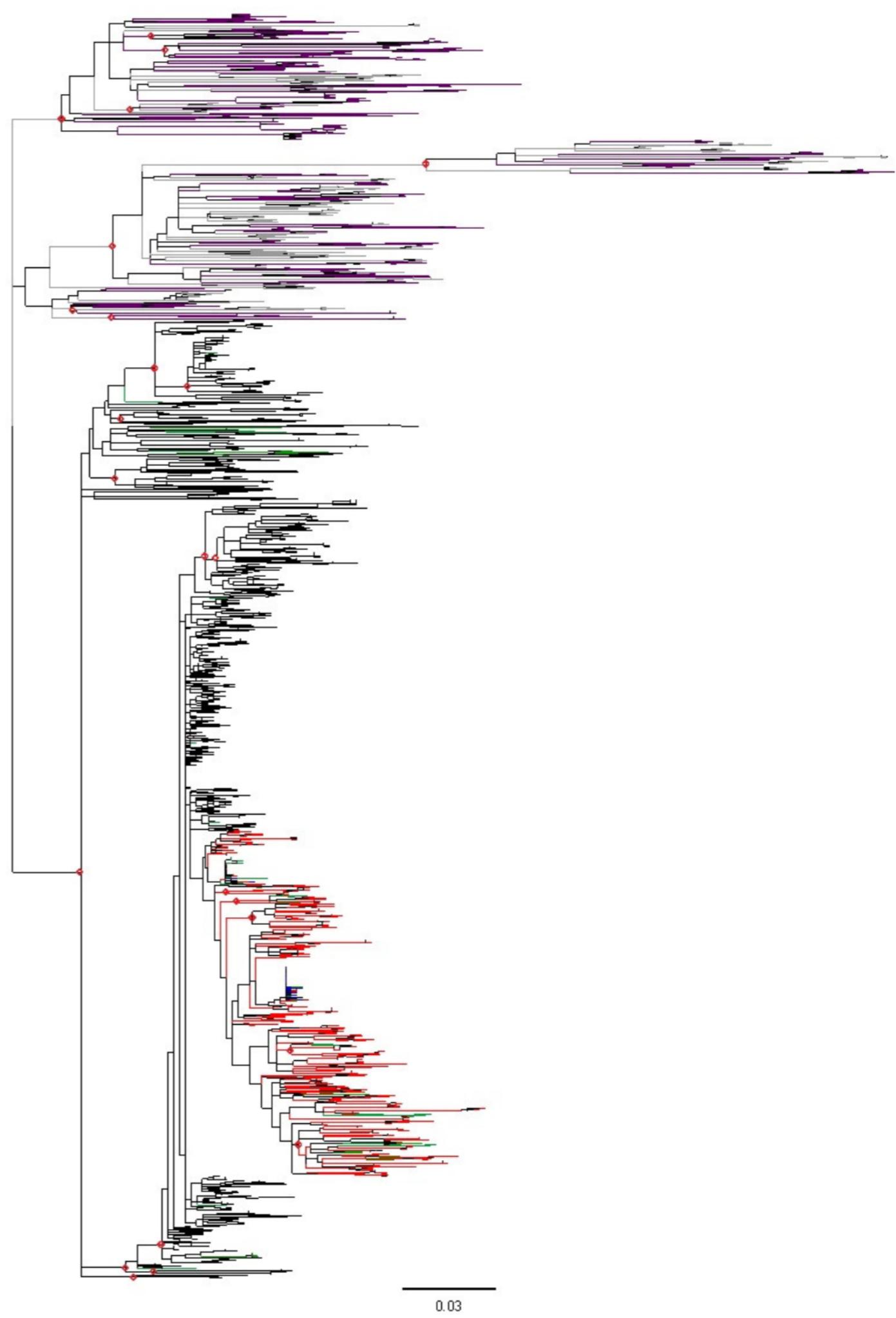
- 4 Pathogen diversity can also be used to corroborate time of infection using the mutation rate as a molecular clock.

# Filogenias como evidencia forense

- Juan Maeso, trabajador de un hospital en valencia
- Acusado de infectar a al menos 275 personas con hepatitis C
- Se inyectaba la morfina de los pacientes y después, con la misma jeringa, inyectaba a los pacientes
- 4 muertos
- 20 años preso



- Púrpura = sin relación al brote
- Gris = controles locales
- Negro, rojo y verde = grupo de pacientes y supuesto culpable



# Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source

Fernando González-Candelas  , María Alma Bracho, Borys Wróbel and Andrés Moya

BMC Biology 2013 11:76 | DOI: 10.1186/1741-7007-11-76 | © González-Candelas et al.; licensee BioMed Central Ltd. 2013

Received: 26 July 2012 | Accepted: 24 May 2013 | Published: 19 July 2013

COMMENTARY | OPEN ACCESS

# Viral phylogeny in court: the unusual case of the Valencian anesthetist

Anne-Mieke Vandamme  and Oliver G Pybus 

BMC Biology 2013 11:83 | DOI: 10.1186/1741-7007-11-83 | © Vandamme and Pybus; licensee BioMed Central Ltd. 2013

Received: 1 July 2013 | Accepted: 11 July 2013 | Published: 19 July 2013

NATURE | NEWS FEATURE

عربي



## Science in court: Disease detectives

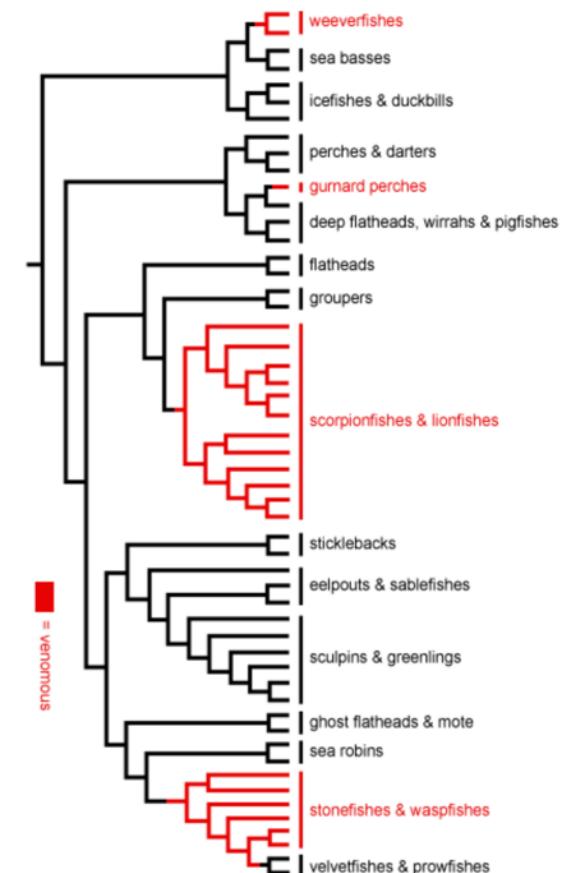
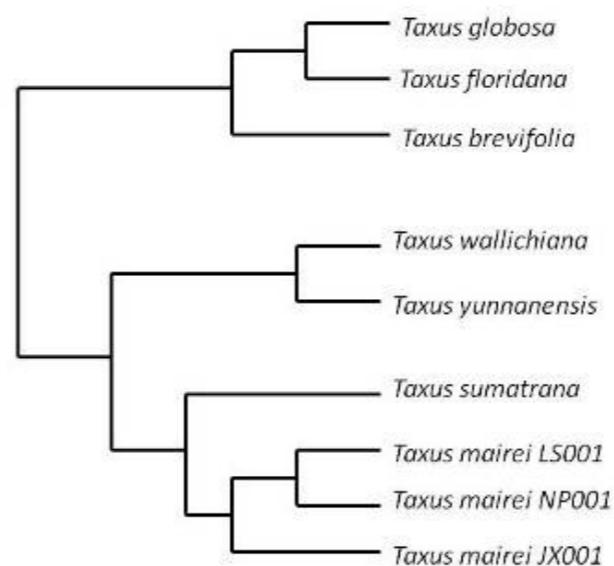
A powerful method for deducing microbial relationships has been edging its way into civil and criminal investigations. But courts should proceed with caution.

Shaoni Bhattacharya

26 February 2014

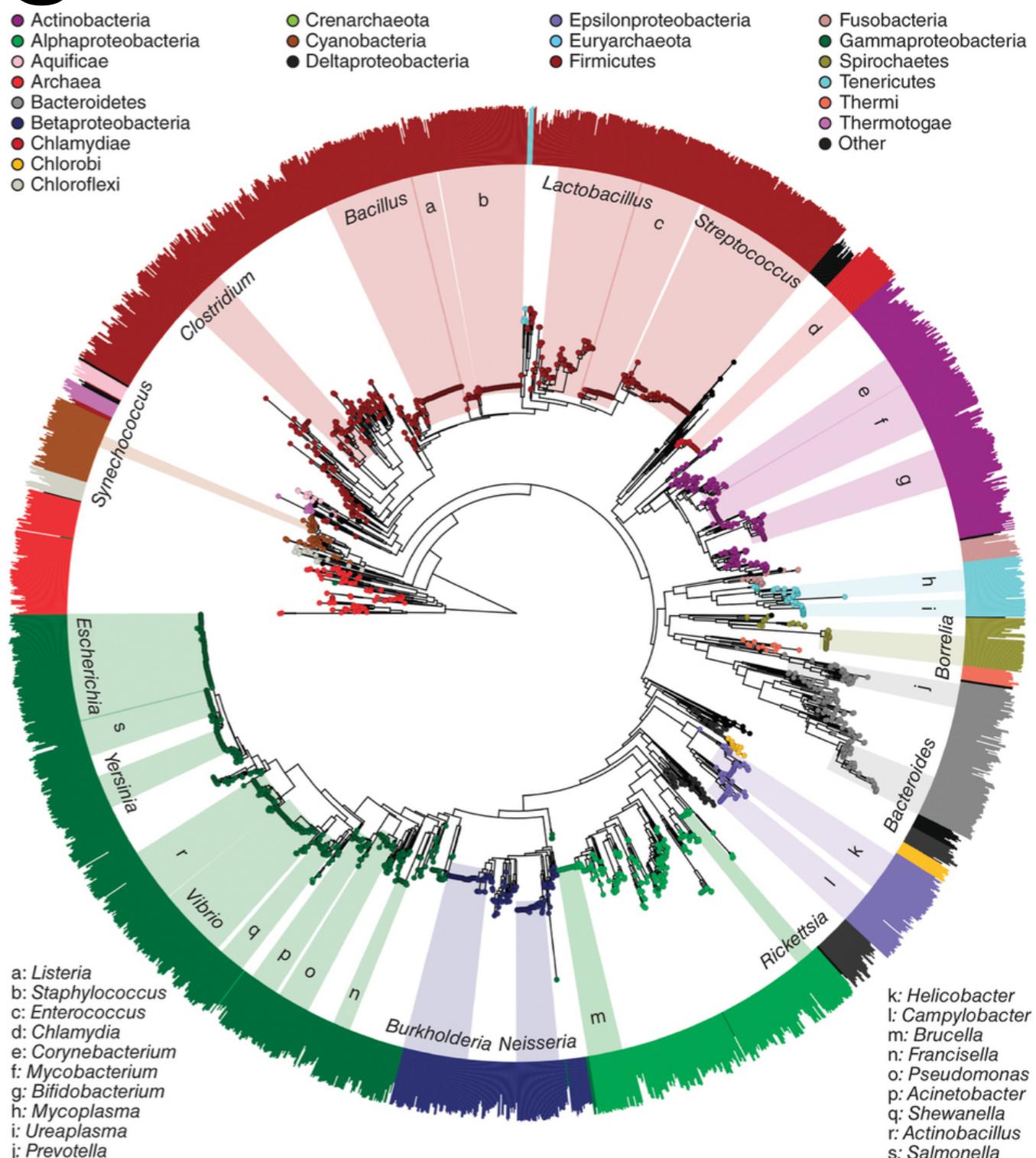
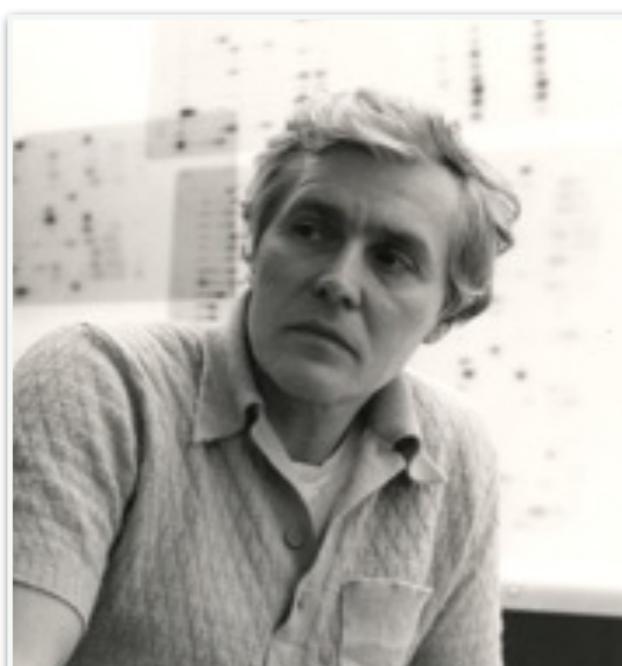
# Ejemplo sobre la utilidad de las filogenias

- Búsqueda de compuestos naturales = drug discovery
- Es más probable encontrar compuestos similares en organismos relacionados evolutivamente
- Ejemplo = taxol = droga contra el cáncer = agente antimicrotúbulos



# Ejemplo sobre la utilidad de las filogenias

- Identificar especies y comunidades bacterianas a través de secuenciación del gen 16S rRNA
- Propuesto por Carl Woese en 1977

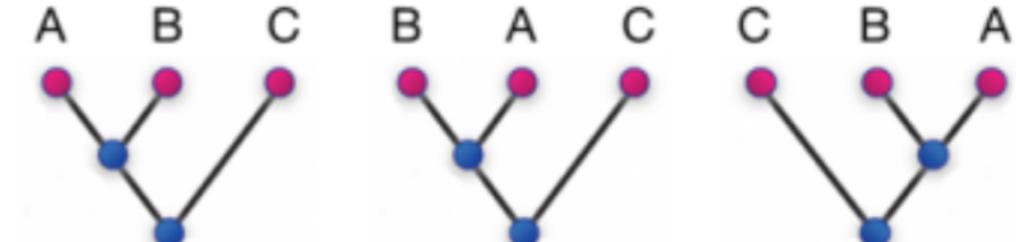


Algunas definiciones y  
características de los  
árboles filogenéticos

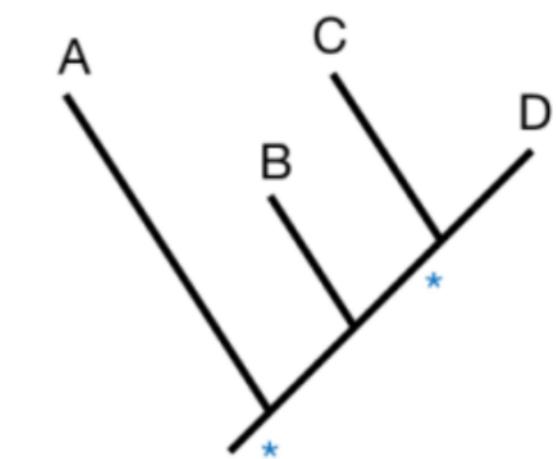
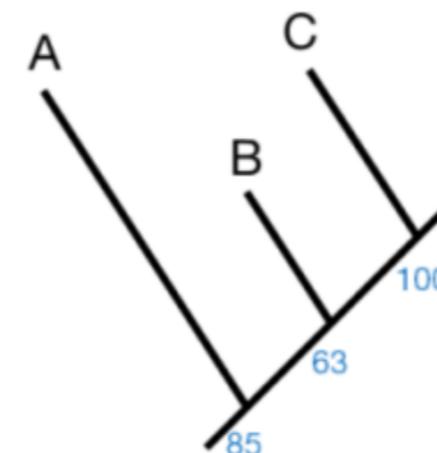
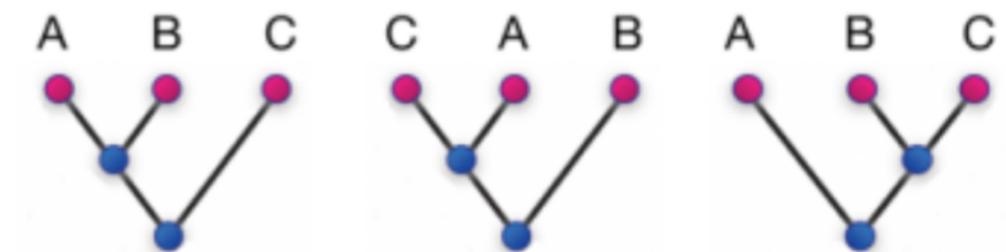
# Características de las filogenias

- Topología
- Ramas - branches (edges)
- Nodos - nodes (vertices)
  - Raíz - root
  - Nodos internos
  - Extremos - tips
- Confianza en nodos

These trees display the same topology

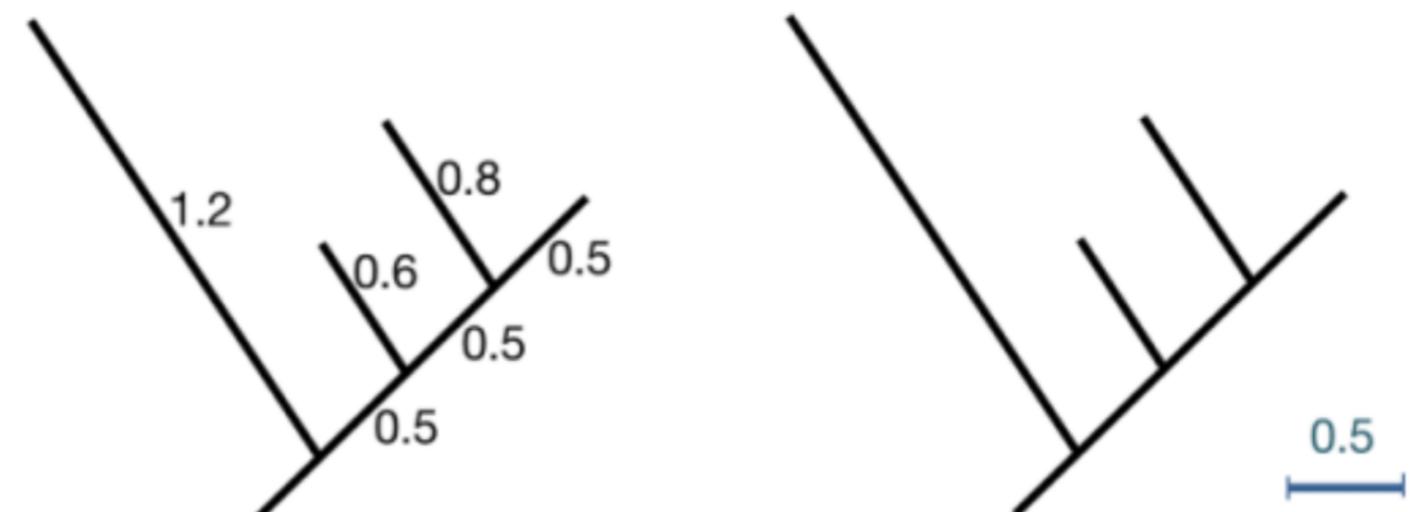


These trees display different topologies



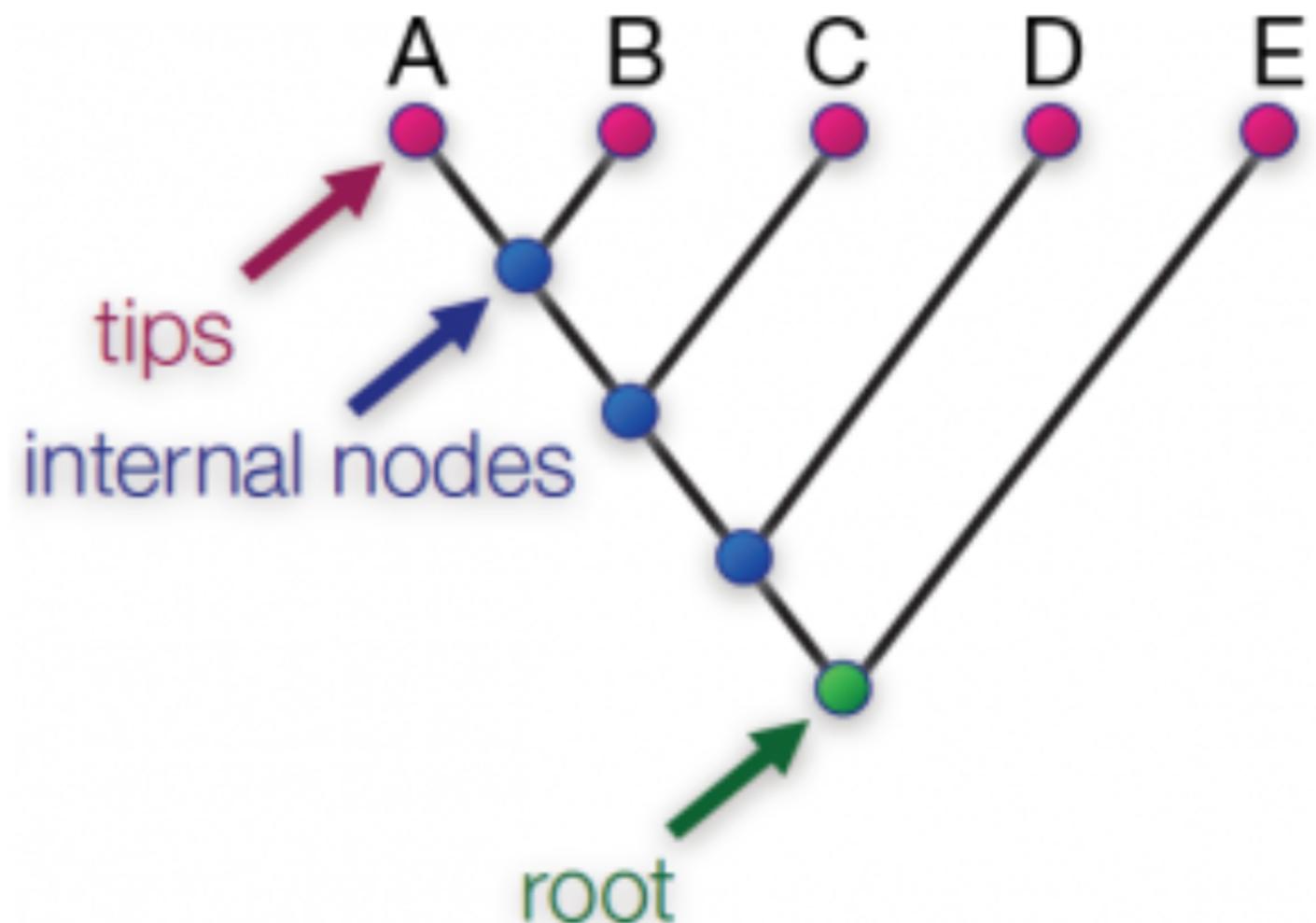
# Características de las filogenias

- Ramas - branches (edges)
- Indican la cantidad de cambio
- Expresado en sustituciones por sitio, sustituciones por gen
- En algunos casos, el largo de las ramas está escalado por tiempo (relojes moleculares)



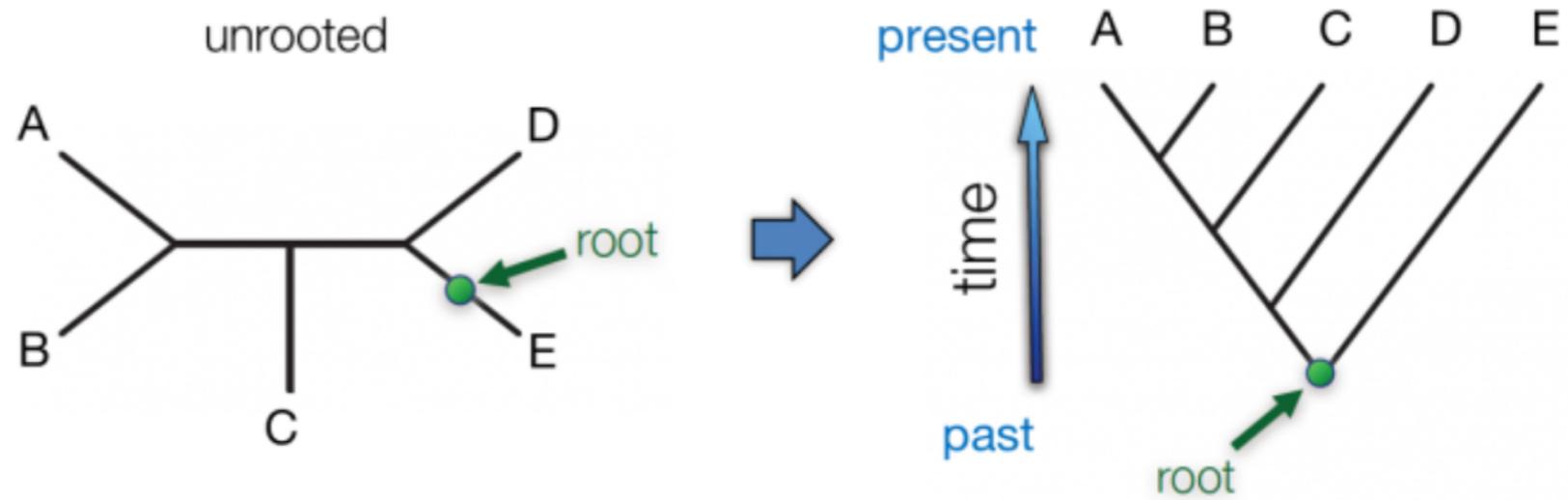
# Características de las filogenias

- Nodos
- Representan ancestros
- Mayoría de las veces no podemos observar a esos ancetros
- Excepciones?

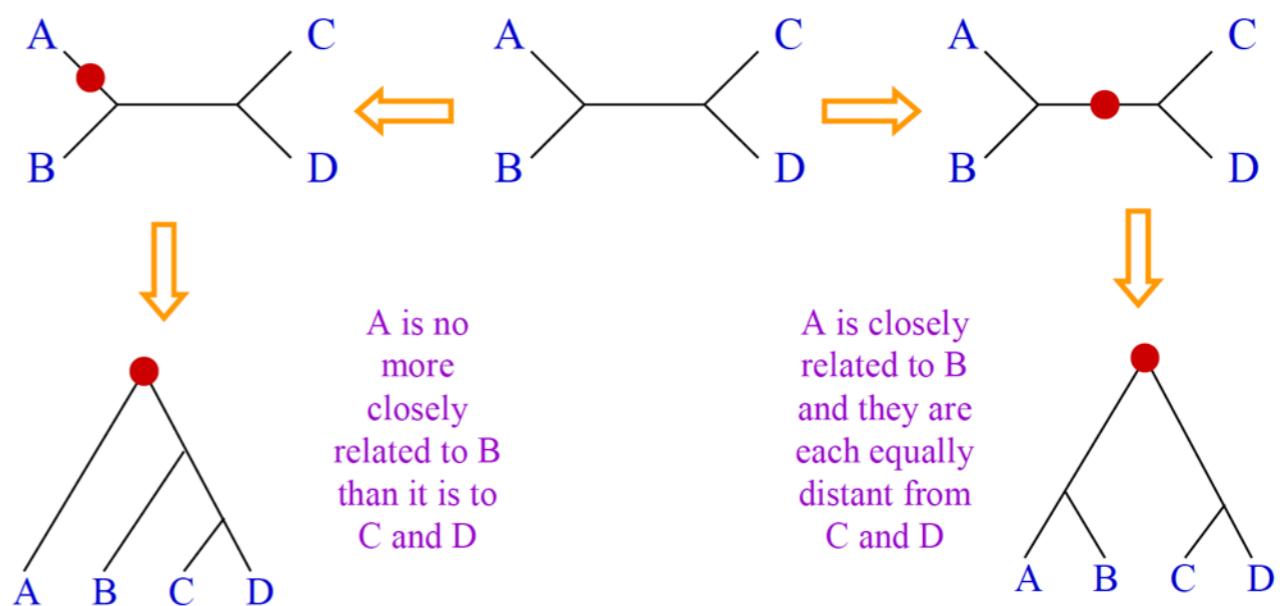


# Características de las filogenias

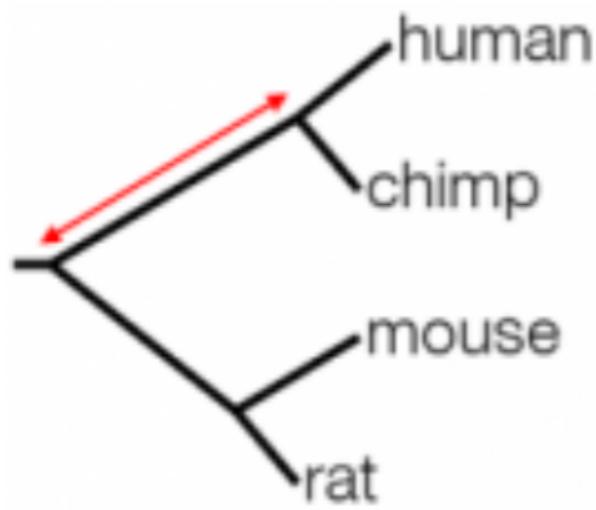
- Raíz
- Árboles enraizados y no enraizados
- Posibles árboles?
- Enraizados =  $(2n-3)!!$
- No-enraizados =  $(2n-5)!!$



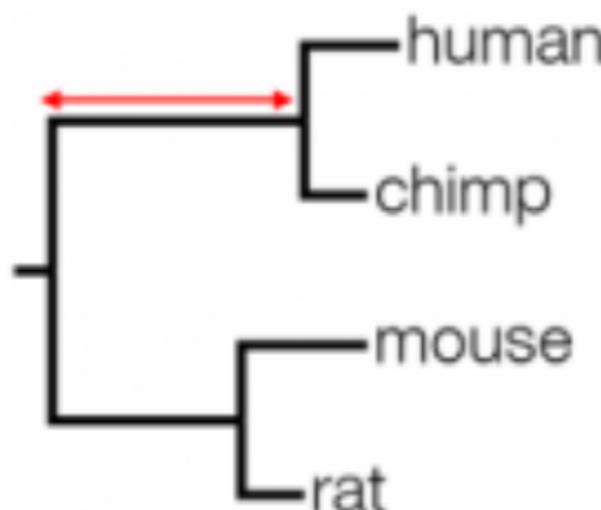
- A single unrooted tree can imply different relationships between species depending on the location of the root



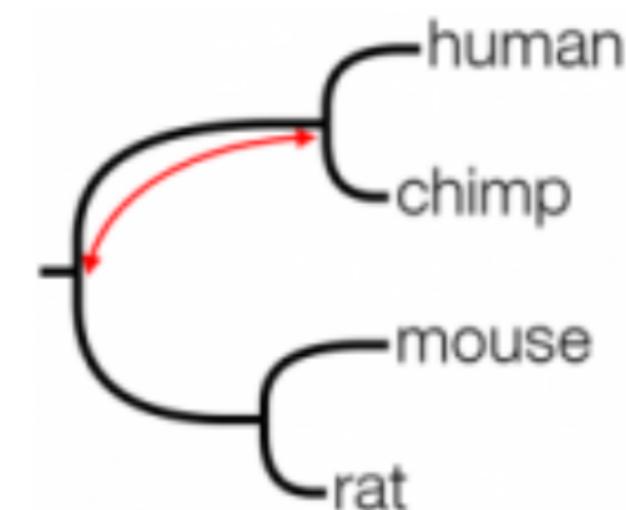
# Características de las filogenias



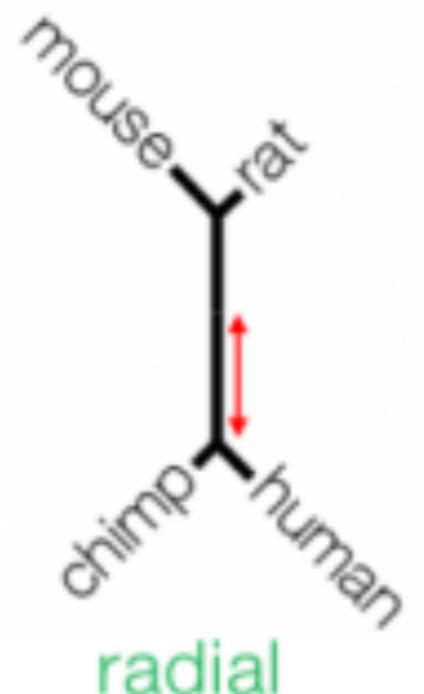
diagonal



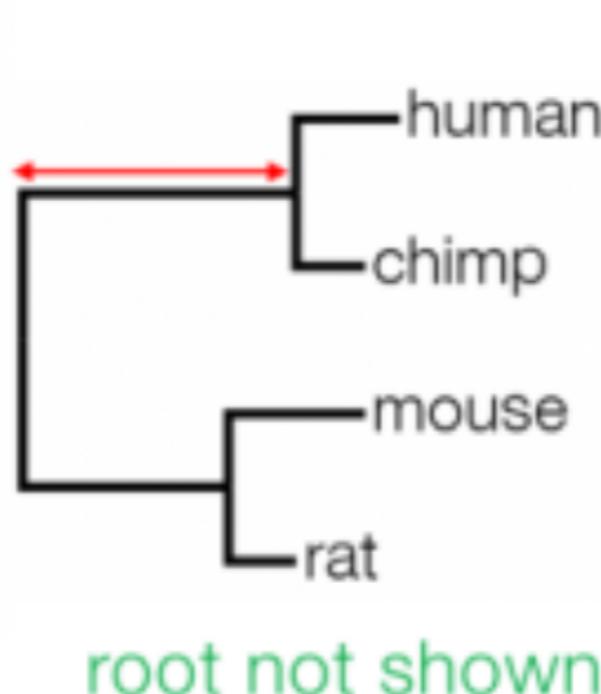
rectangular



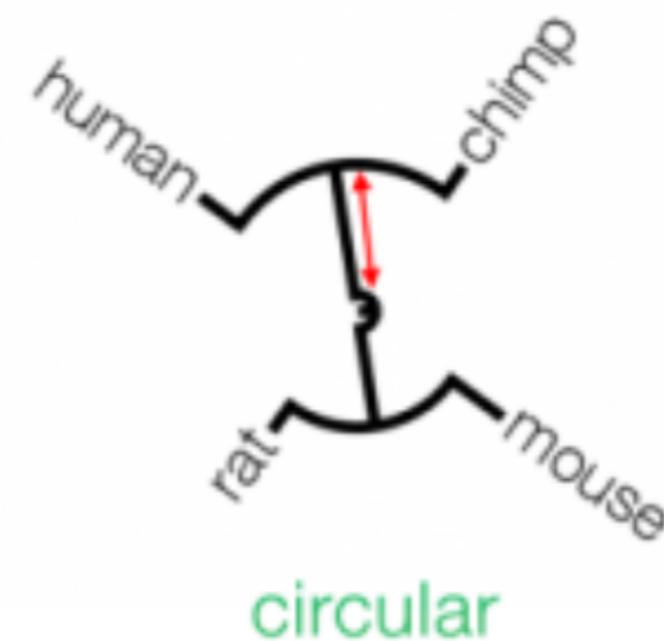
curved



radial



root not shown

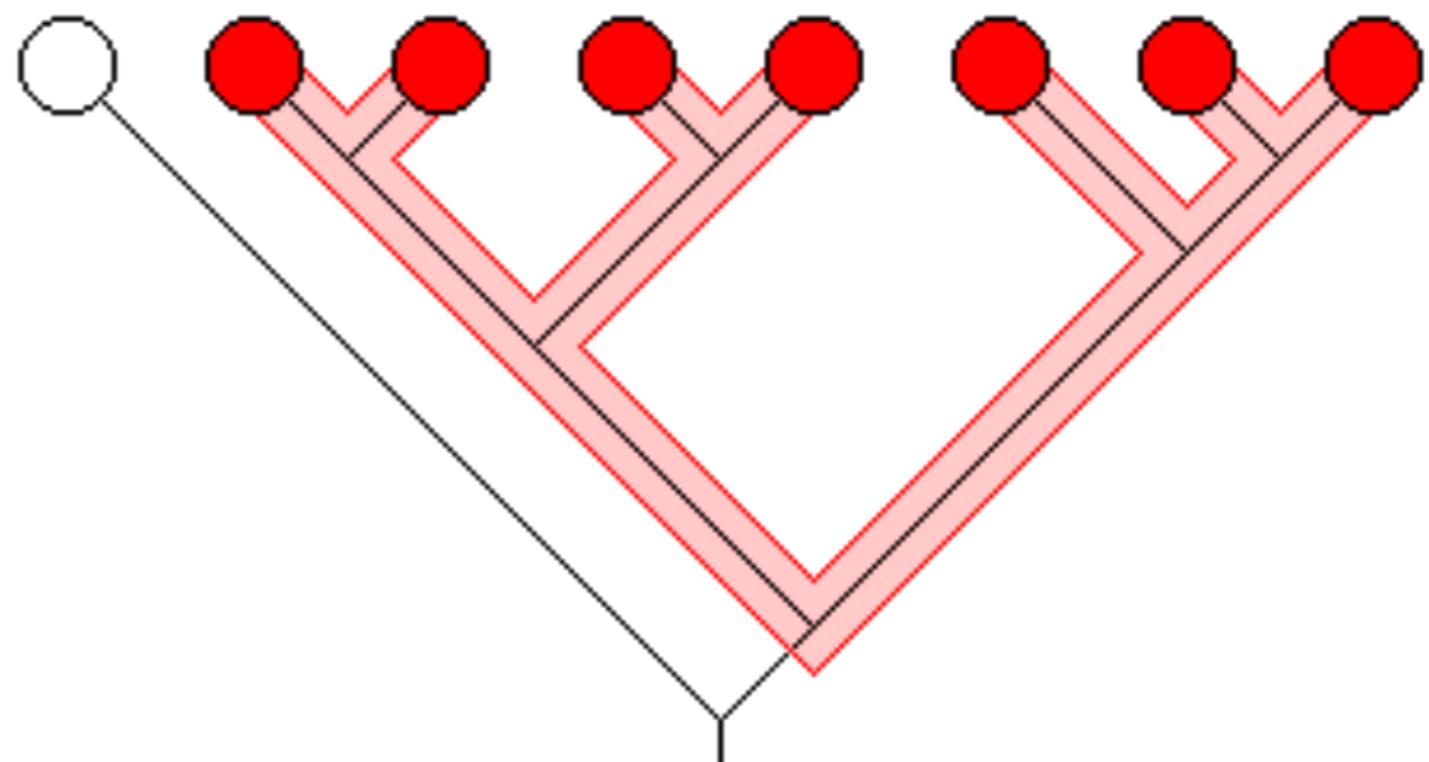


circular

# Biología en filogenias

- Grupo monofilético o clado
- Un grupo compuesto de una colección de organismos que comprende un ancestro y todos sus descendientes
- Mamíferos, aves, plantas angioespermas, etc.

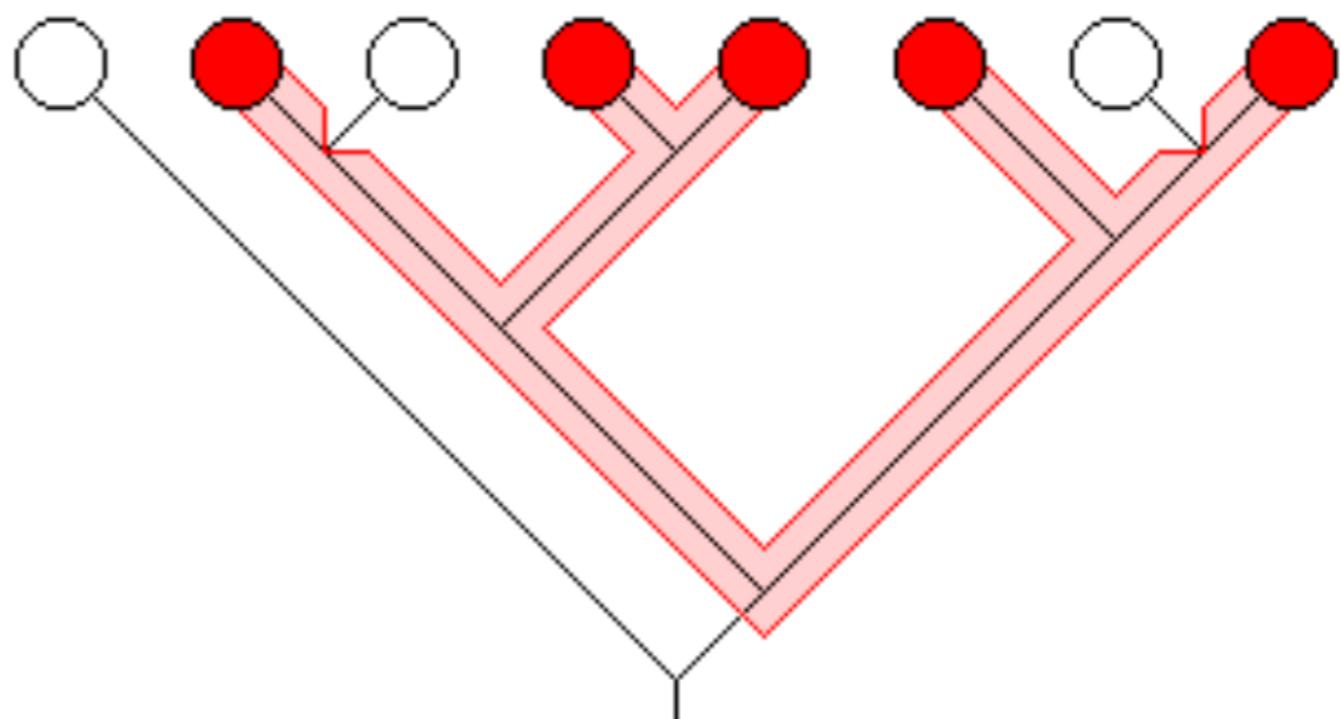
**Monophyletic taxon (clade) :**



# Biología en filogenias

- Grupo parafilético
- Un grupo compuesto de una colección de organismos que comprende un ancestro pero NO a todos sus descendientes
- Peces, gimnoespermas, protistas

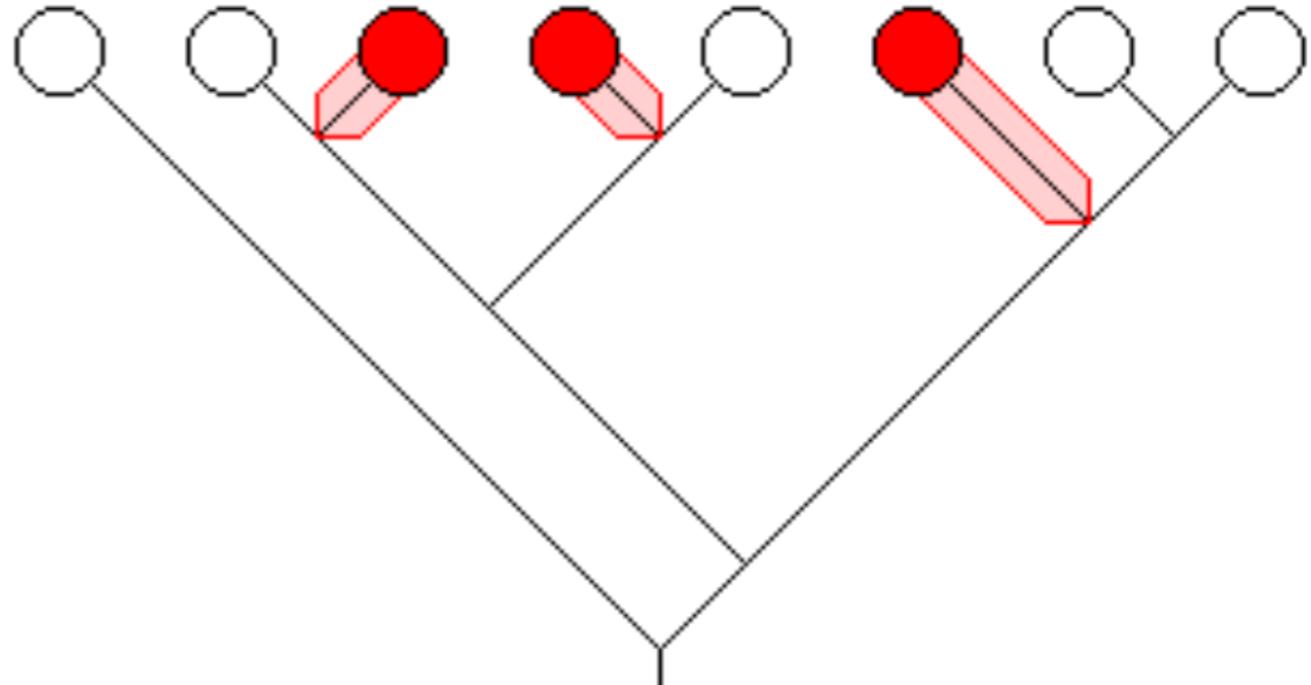
**Paraphyletic taxon :**



# Biología en filogenias

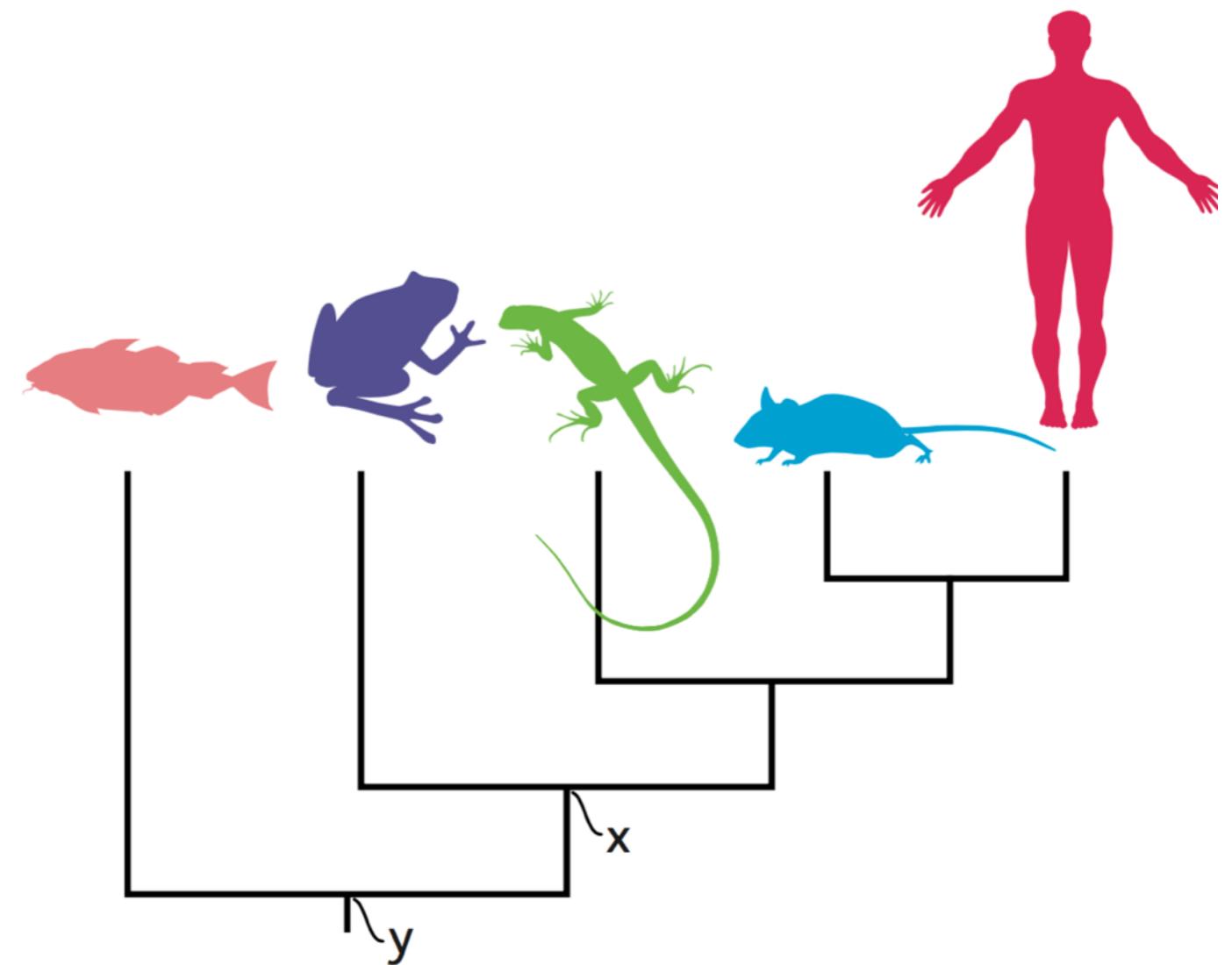
- Grupo polifilético
- Un grupo compuesto de una colección de organismos donde el ancestro en común de todos ellos no está incluido
- mamíferos marinos, árboles, algas, etc.

**Polyphyletic taxon :**



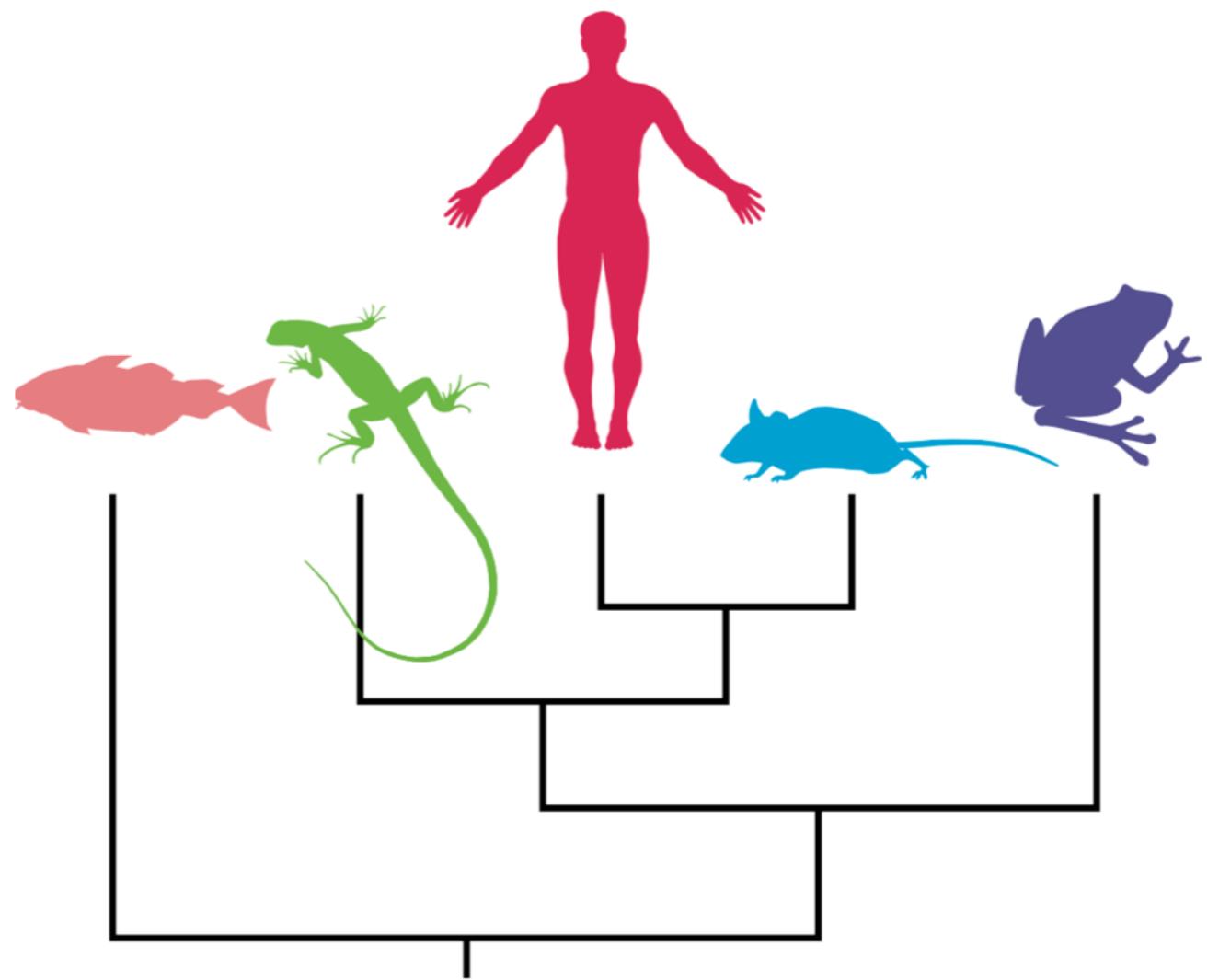
# ¿Cómo “leer” una filogenia?

- Basándose en el árbol ¿la rana es más cercana al pez o al humano?



# ¿Cómo “leer” una filogenia?

- ¿Y ahora?

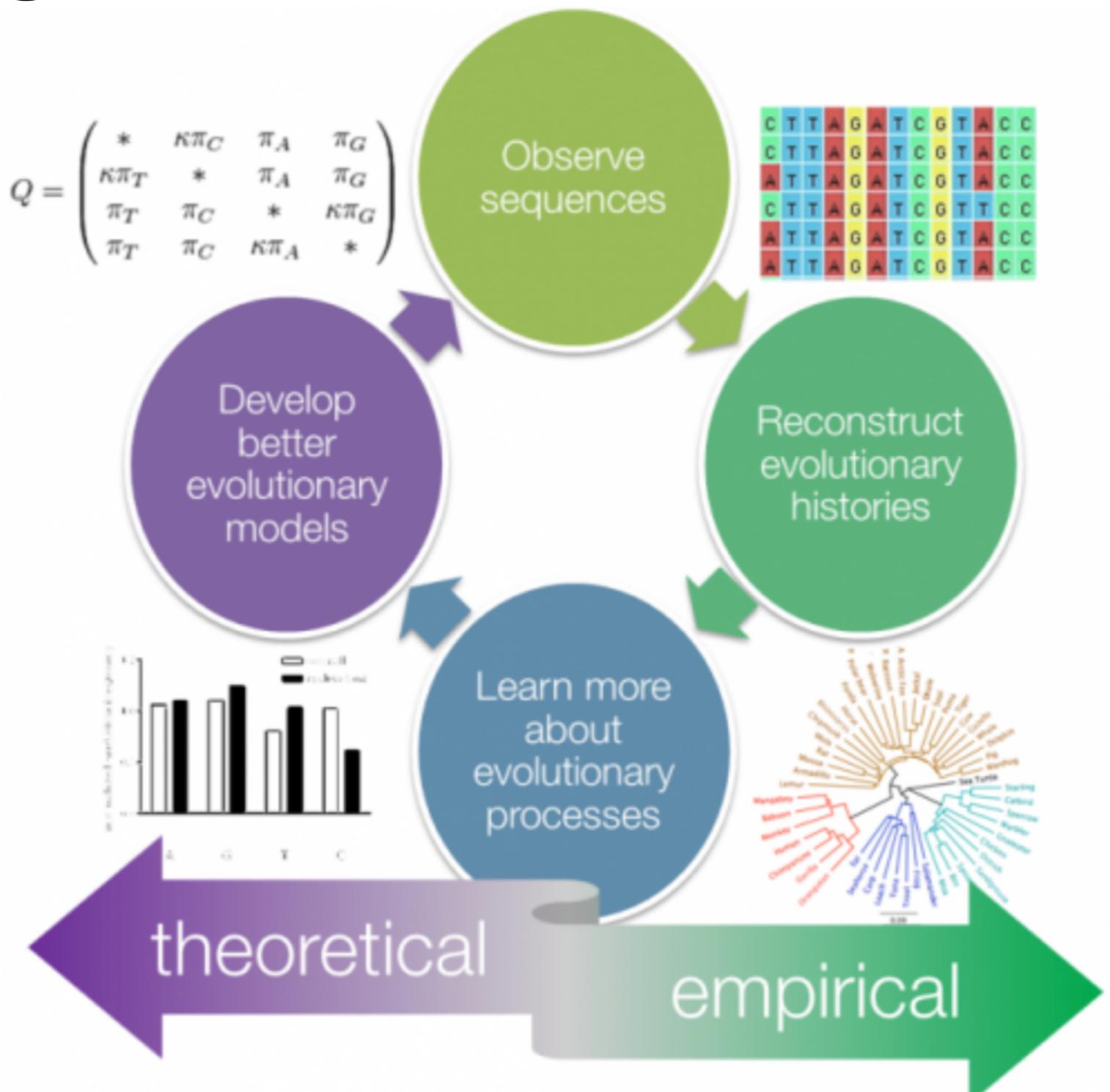


Baño, café, aire fresco  
15 minutos de recreo

¿Cómo inferimos una  
filogenia?

# ¿Cómo se infieren las filogenias?

- Genes, genomas, caracteres morfológicos
- Taxa, taxon (plural, singular)

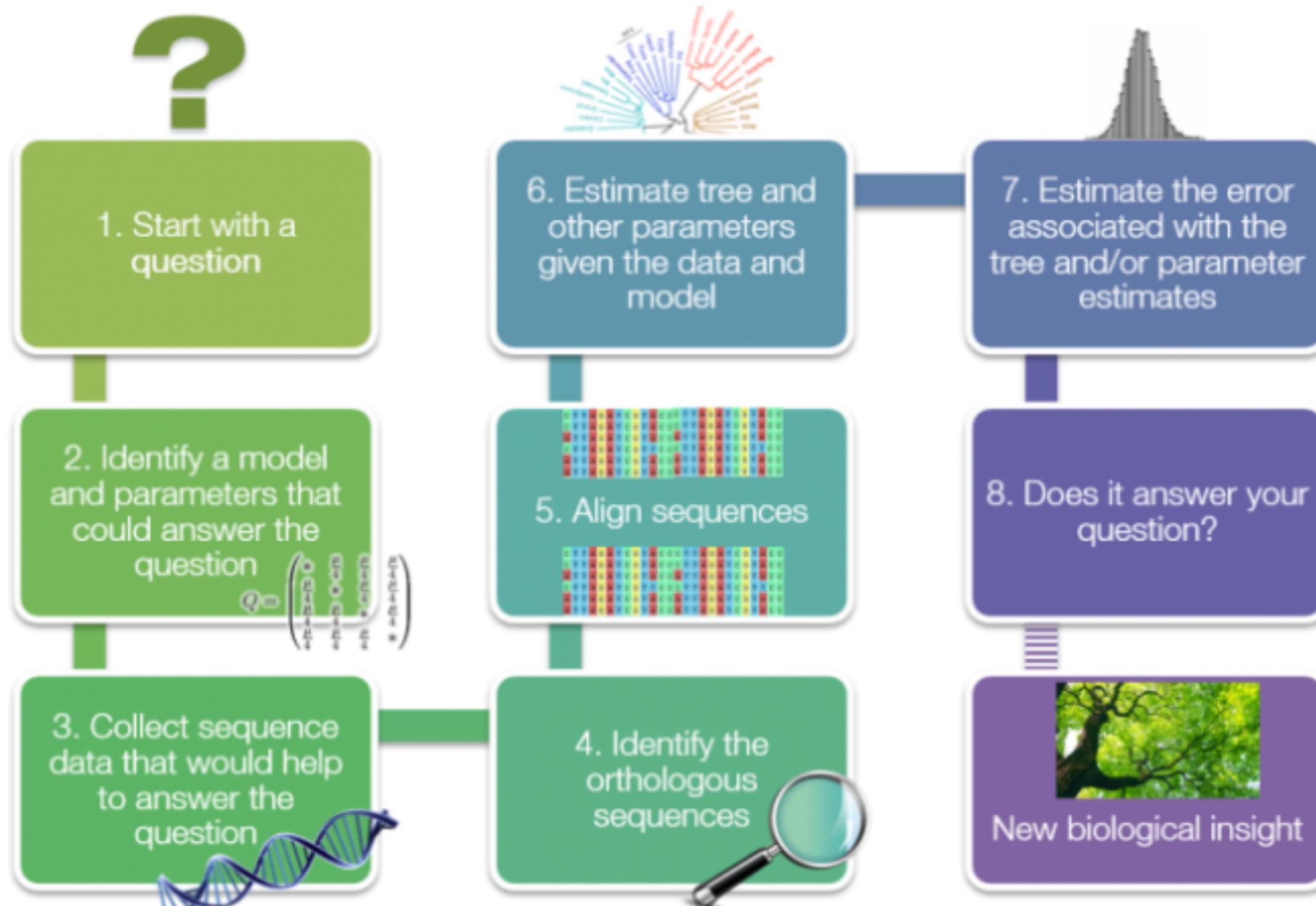


# ¿Cómo se infieren las filogenias?

- Alineamiento múltiple
- Modelo de sustitución nucleotídica
- Calcular filogenia
- No es posible hacerlo de manera exacta

n	B <sub>n</sub>	b' <sub>n</sub>
3	1	3
4	3	15
5	15	105
6	105	945
7	954	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425
20	2.22E+020	8.20E+021
30	8.69E+036	4.95E+038

# ¿Cómo se infieren las filogenias?



# Modelos de sustitución nucleotídica

- Modelos que intentan modelar la cantidad de cambio en un grupo de secuencias

- Manera más simple:

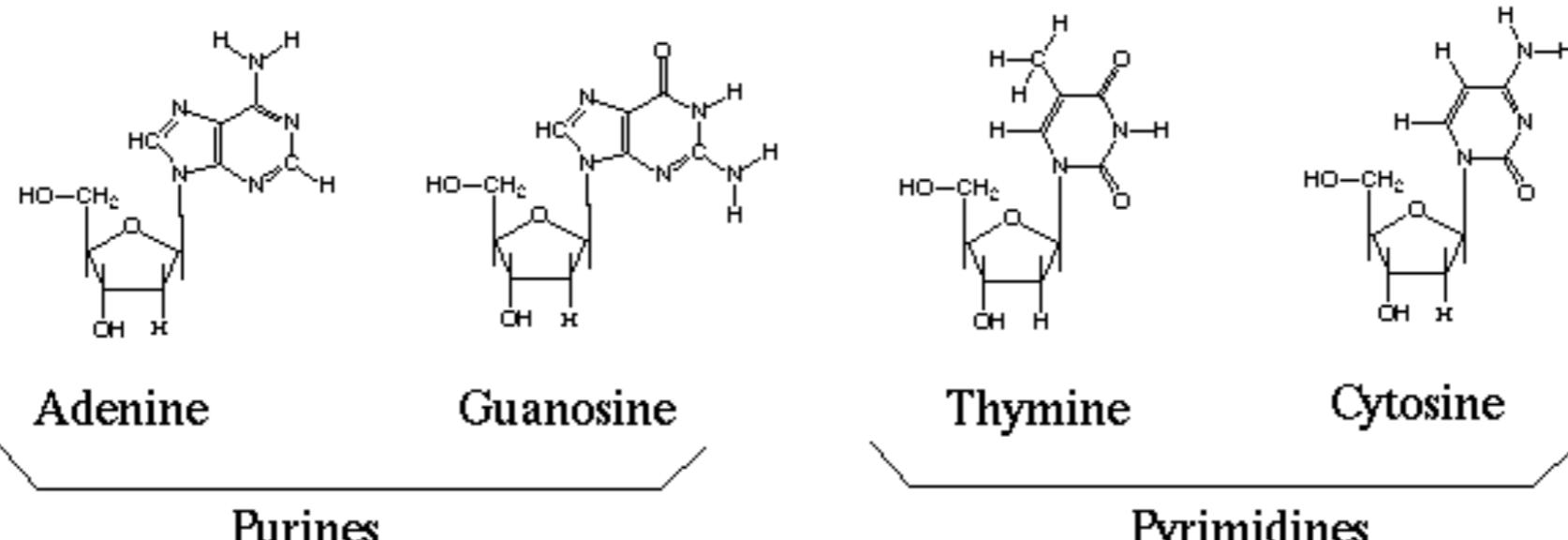
Human **ATG**T**TGACTC**

Mouse **ATG**C**TGACTC**

- 1/10, 0.1 sustituciones por sitio
- Esto asume que hemos observado todas las sustituciones que han ocurrido
- También asumimos que todas los cambios, T → C y A → G ocurren con la misma frecuencia

# Modelos de sustitución nucleotídica

## The Nucleotides of DNA



### Sequence of part of a normal gene

### Sequence of mutated gene

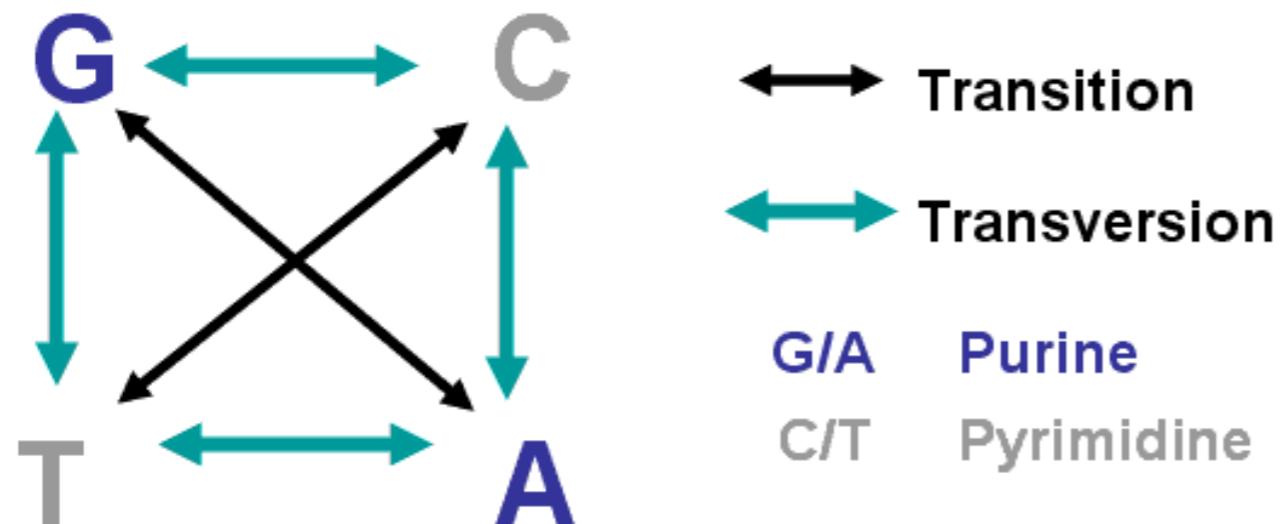
a) Transition mutation (A-T to G-C in this example)

DNA 5' TCTCAAAAATTTACG 3' 3' AGAGTTTTAAATGC 5'	5' TCTCAAGAATTTACG 3' 3' AGAGTTCTTAAATGC 5'
---	--

**b) Transversion mutation (C-G to G-C in this example)**

$5'$ TCT <b>C</b> AAAAAATTTACG $3'$ AGAGTTTTAAATGC	$3'$ TCT <b>G</b> AAAAAATTTACG $5'$ AGAC <b>T</b> TTTTAAATGC
---	---

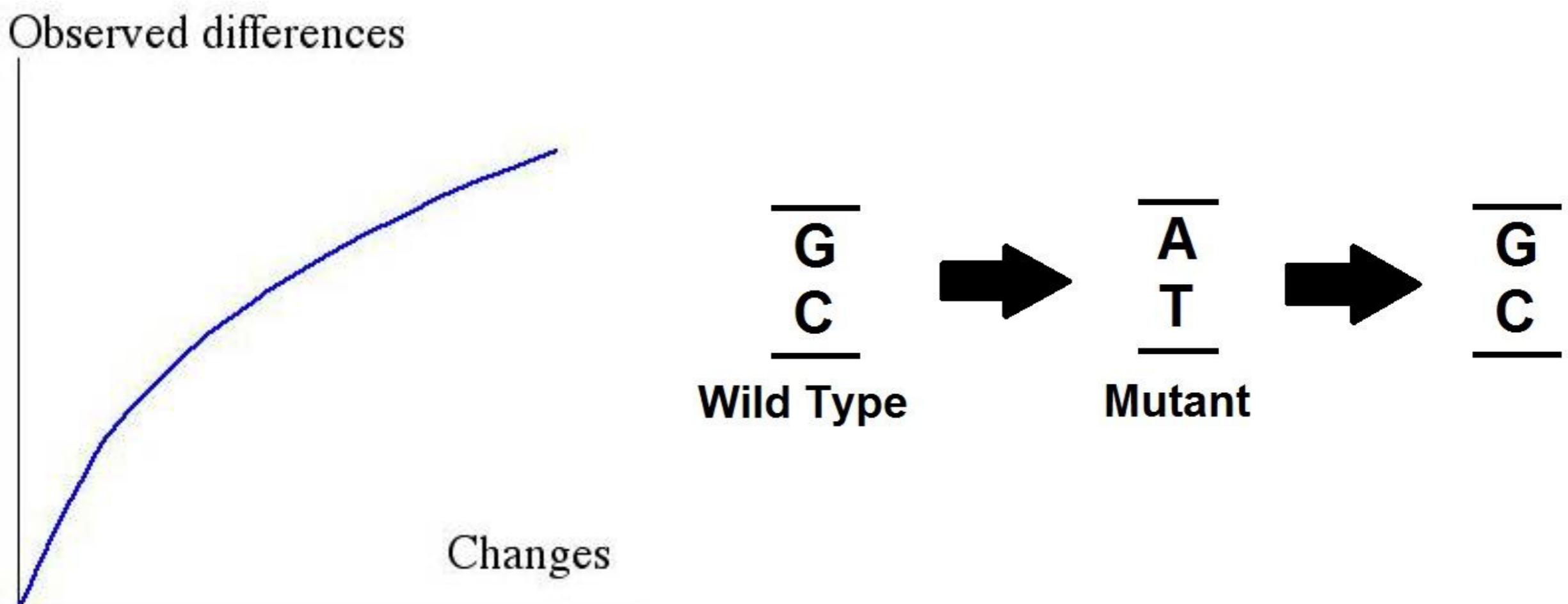
# Modelos de sustitución nucleotídica



- Transiciones son más probables que las transversiones

# Modelos de sustitución nucleotídica

- Mutaciones múltiples en la misma posición
- Multiple hits and back mutations



# Modelos de sustitución nucleotídica

- Jukes-Cantor = más simple, todas las tasas de sustitución son iguales, frecuencia de bases iguales
- Kimura 1980 (K80) = Distingue transiciones y transversiones
- Felsenstein 1981 (F81) = Incluye frecuencias desiguales
- Hasegawa, Kishino and Yano 1985 (HKY) = Distingue transiciones y transversiones, además de incluir frecuencias desiguales

# Modelos de sustitución nucleotídica

- Generalized time-reversible (Tavaré 1986) o simplemente GTR
- Tasas de sustitución varian, la frecuencia de bases es desigual

$$\begin{array}{ccccc} & A & G & C & T \\ \begin{matrix} A \\ G \\ C \\ T \end{matrix} & \left( \begin{matrix} - & \pi_G rAG & \pi_C rAC & \pi_T rAT \\ \pi_A rAG & - & \pi_C rCG & \pi_T rGT \\ \pi_A rAC & \pi_G rCG & - & \pi_T rCT \\ \pi_A rAT & \pi_G rGT & \pi_C rCT & - \end{matrix} \right) \end{array}$$

# ¿Cómo escoger el modelo correcto para mis datos?

- Siempre escoger el modelo más cercano a tus genes
- Usando criterios estadísticos = Akaike Information Criterion, Bayesian Information Criterion, Likelihood Ratio Test, etc.
- ModelTest program

# ¿Cómo escoger el modelo correcto para mis datos?

- ModelTest program → 76 paper más citado de la historia de la ciencia

Modeltest: testing the model of DNA substitution.

Authors David Posada, Keith A. Crandall

Publication date 1998/1/1

Journal Bioinformatics

Volume 14

Issue 9

Pages 817-818

Publisher Oxford University Press

Description SUMMARY: The program MODELTEST uses log likelihood scores to establish the model of DNA evolution that best fits the data. AVAILABILITY: The MODELTEST package, including the source code and some documentation is available at [http://bioag.byu.edu/zoology/crandall\\_lab/modeltest.html](http://bioag.byu.edu/zoology/crandall_lab/modeltest.html).

Total citations [Cited by 18228](#)

NATURE | NEWS FEATURE

عربي

## The top 100 papers

**Nature explores the most-cited research of all time.**

**Richard Van Noorden, Brendan Maher & Regina Nuzzo**

29 October 2014

# ¿Cómo inferir filogenias?

- Caracteres o distancia
- Criterios de optimalidad o deterministas

# Métodos de distancia

- Son métodos rápidos, tomados de “clustering” en estadística
- Neighbor-joining, Unweighted Pair Group Method with Arithmetic Mean (UPGMA), Least Squares
- Muy rápidos pero muy limitados, se pierde la información nucleotídica, fallan con tasas desiguales de sustitución, etc.
- Muy populares —> investigadores los usan ciegamente

# Criterios de optimalidad

- Producen una distribución de árboles que son ordenados por un criterio
- Utilizan un criterio objetivo para determinar cuál es el mejor árbol dentro de una muestra de árboles
- El más parsimonioso, el con máxima verosimilitud, etc.

# Máxima Parsimonia

- El mejor árbol es el que requiere la menor cantidad de eventos (mutaciones, inserciones, etc.)
- Computacionalmente barato
- Uno de los primeros métodos
- Se ha demostrado que no es estadísticamente consistente

# Máxima Parsimonia

Using Maximum Parsimony  
to Choose Between Two Possible Trees

Sample:

1

2 3

4

5

1

2

3

4

5

Observation:

G

G T

T

G

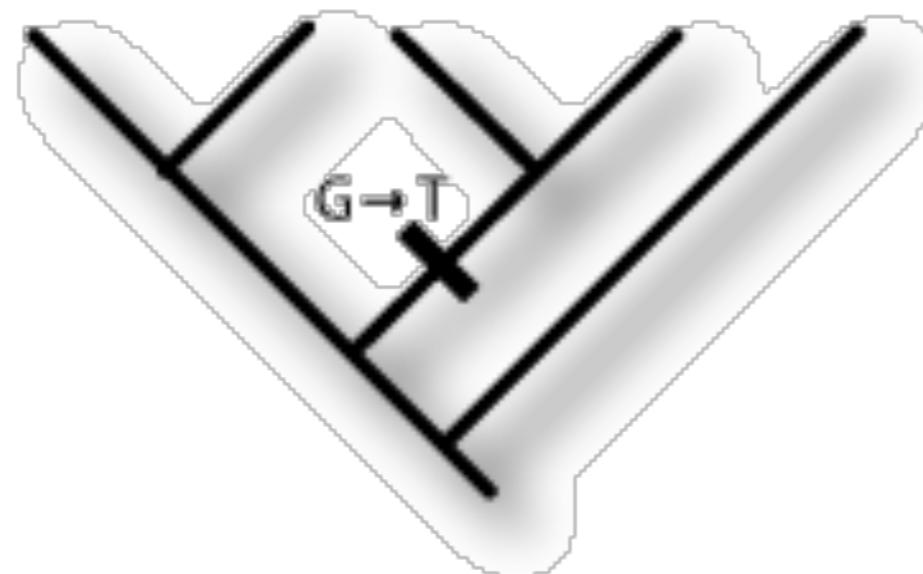
G

G

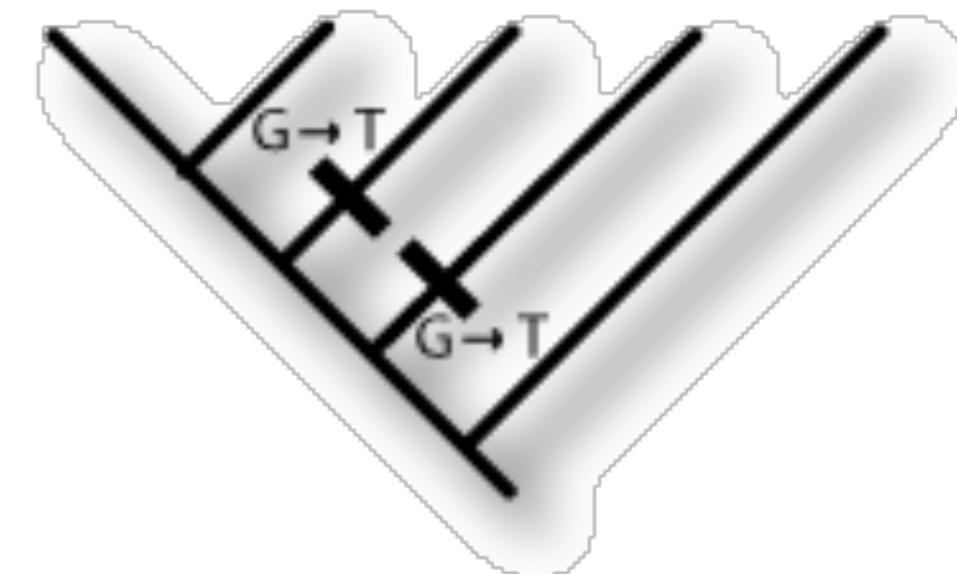
T

T

G



1 change required  
→ better tree



2 changes required  
→ poorer tree

# Máxima Verosimilitud

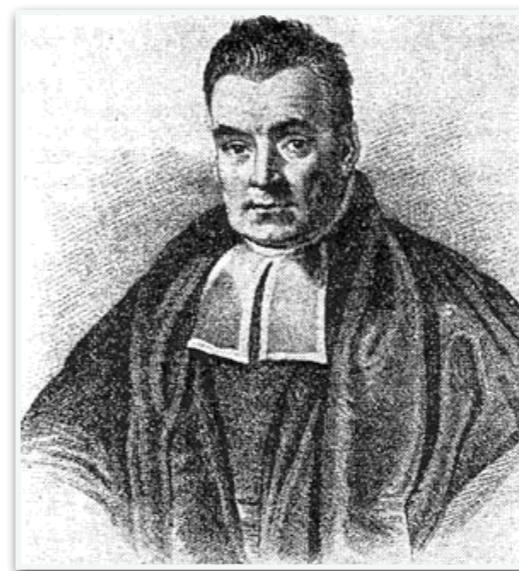
- Maximum Likelihood (ML)
- Si tenemos muchos árboles, ¿cuál árbol se ajusta más a los datos (alineamiento)?
- ¿Cuál es la probabilidad dado el árbol? ¿Cuál árbol tiene la mayor probabilidad de haber generado los datos?
- Buscar muchos posibles árboles y ordenarlos de mayor a menor según valores de ML

# Máxima Verosimilitud

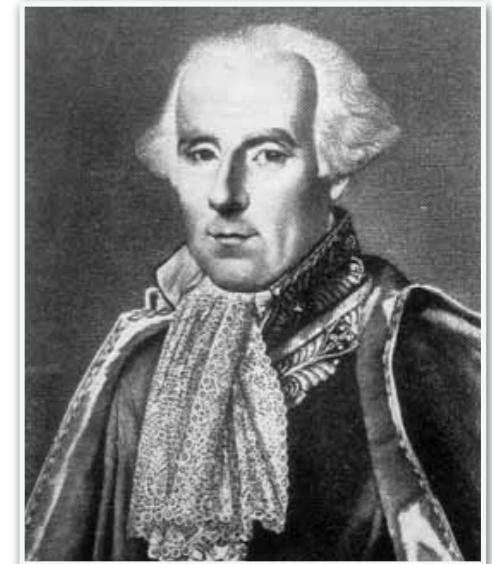
- Maximum Likelihood (ML)
- Computacionalmente más caro
- Robusto a errores de muestreo
- Resultado es dependiente en el modelo de sustitución

# Inferencia Bayesiana

- Usa el teorema de Bayes
- Busca la probabilidad de un árbol según los datos (alineamiento)
- Permite incorporar conocimiento independiente
- Es un método sesgado pero eso está bien



Reverendo  
Thomas Bayes



Pierre Simon  
Laplace

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)},$$

*Bayesian Inference*: The explanation with the highest posterior probability

Bayes' Theorem

$$\Pr(H|D) = \frac{\Pr(H) \Pr(D|H)}{\Pr(D)}$$

Prior probability, the probability of the hypothesis on previous knowledge

Likelihood function, probability of the data given the hypothesis

Posterior probability, the probability of the hypothesis given the data

Unconditional probability of the data, a normalizing constant ensuring the posterior probabilities sum to 1.00

First use in phylogenetics: Li (1996, PhD thesis), Rannala and Yang (1996)

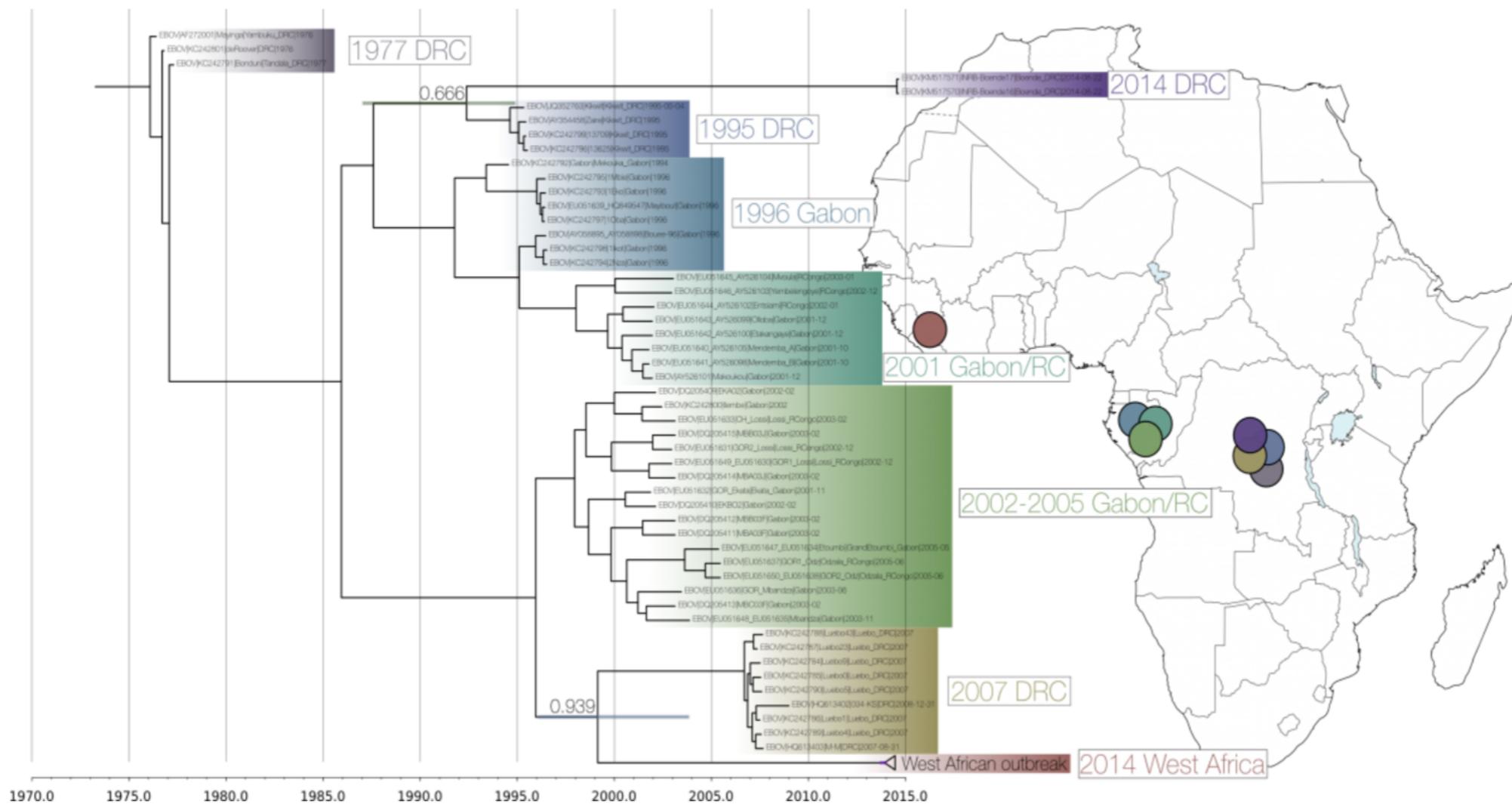
Más allá de la  
filogenia

# Y después de la filogenia?

- Filogeografía
- Reconstrucción de estados ancestrales = vacunas
- Relojes moleculares

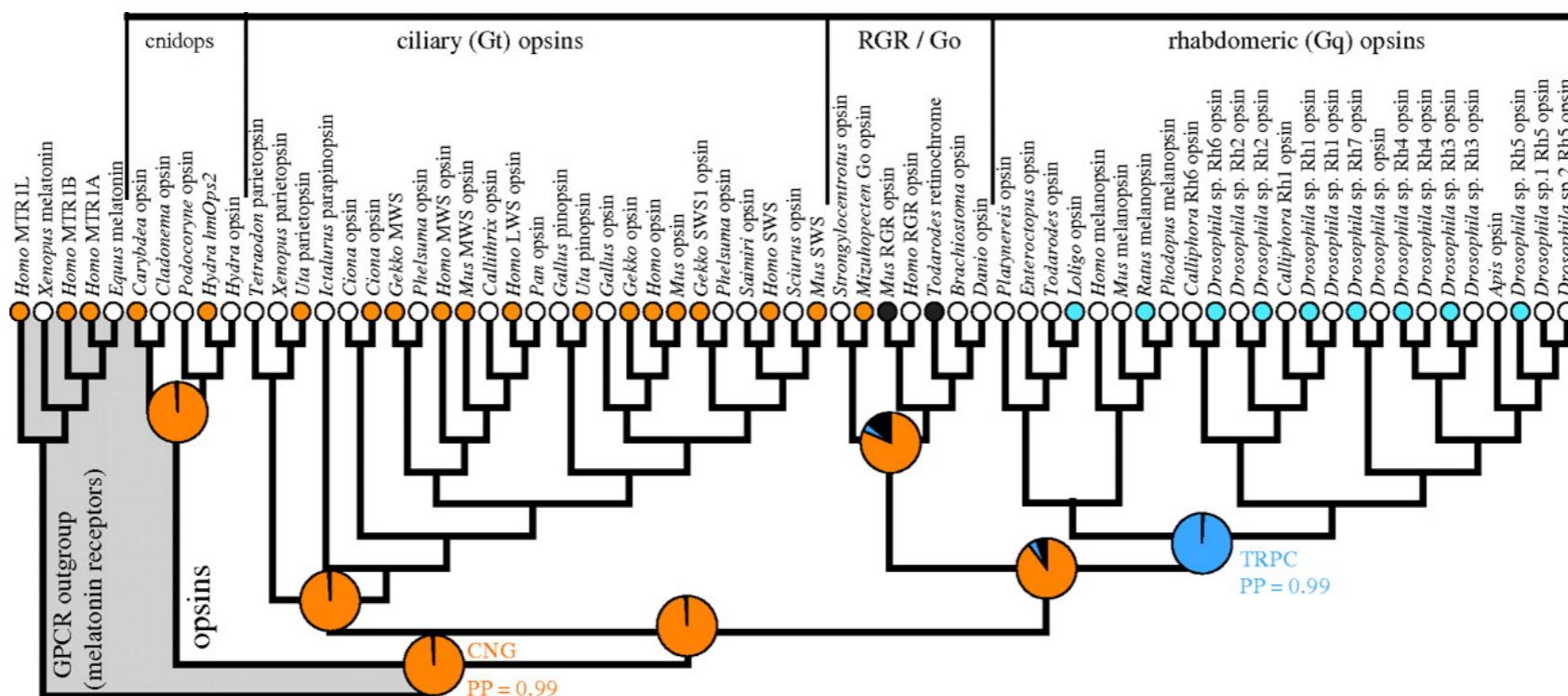
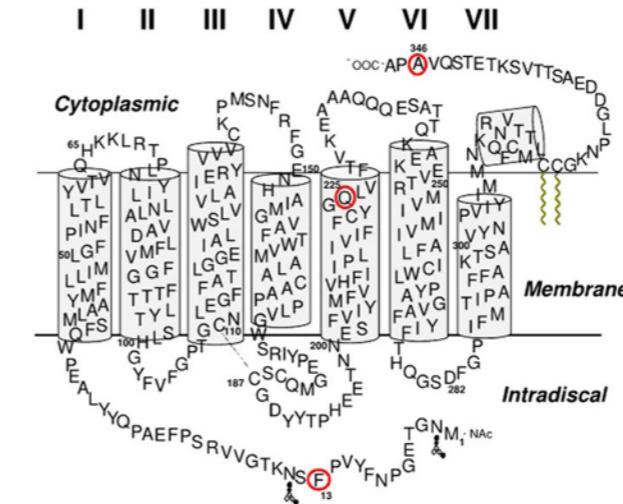
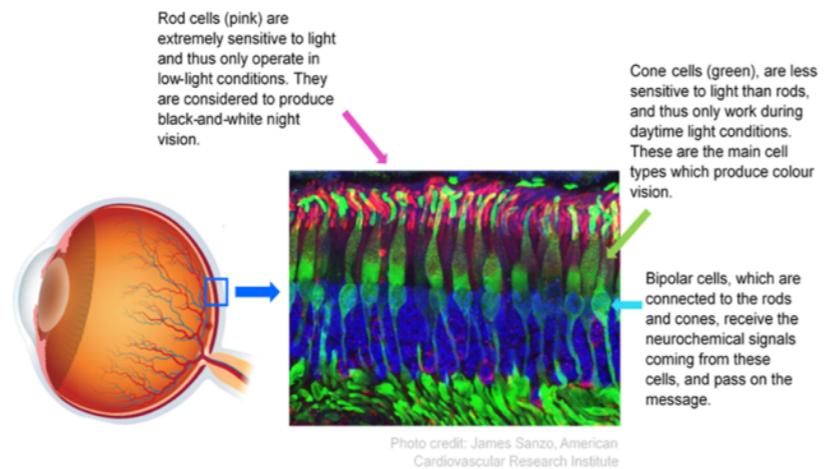
# Filogeografía

- Relacionar historia filogenética y geografía para inferir procesos de dispersión espacial



# Reconstrucción de estados ancestrales

- ¿Cuál es la secuencia ancestral en el nodo x?



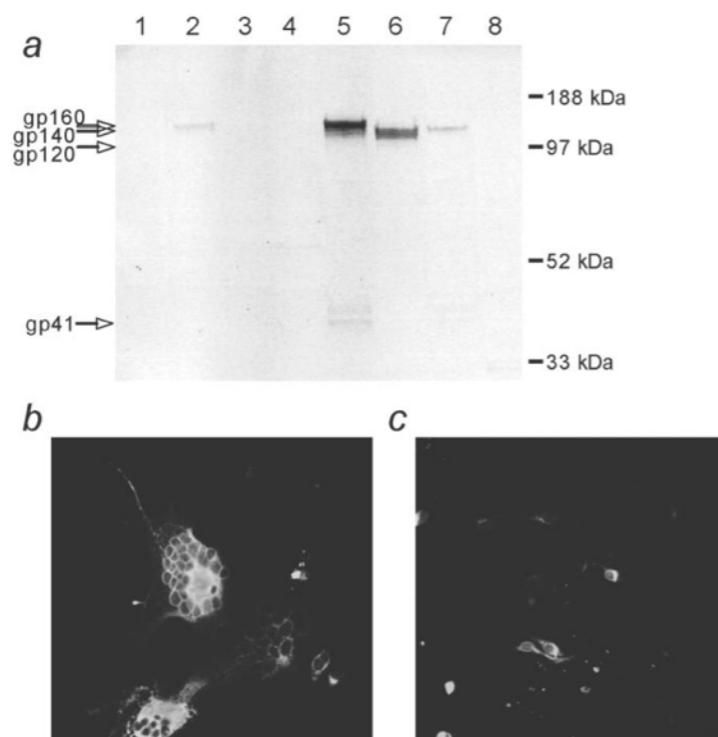
# Reconstrucción de estados ancestrales

- ¿Cuál es la secuencia ancestral en el nodo x?

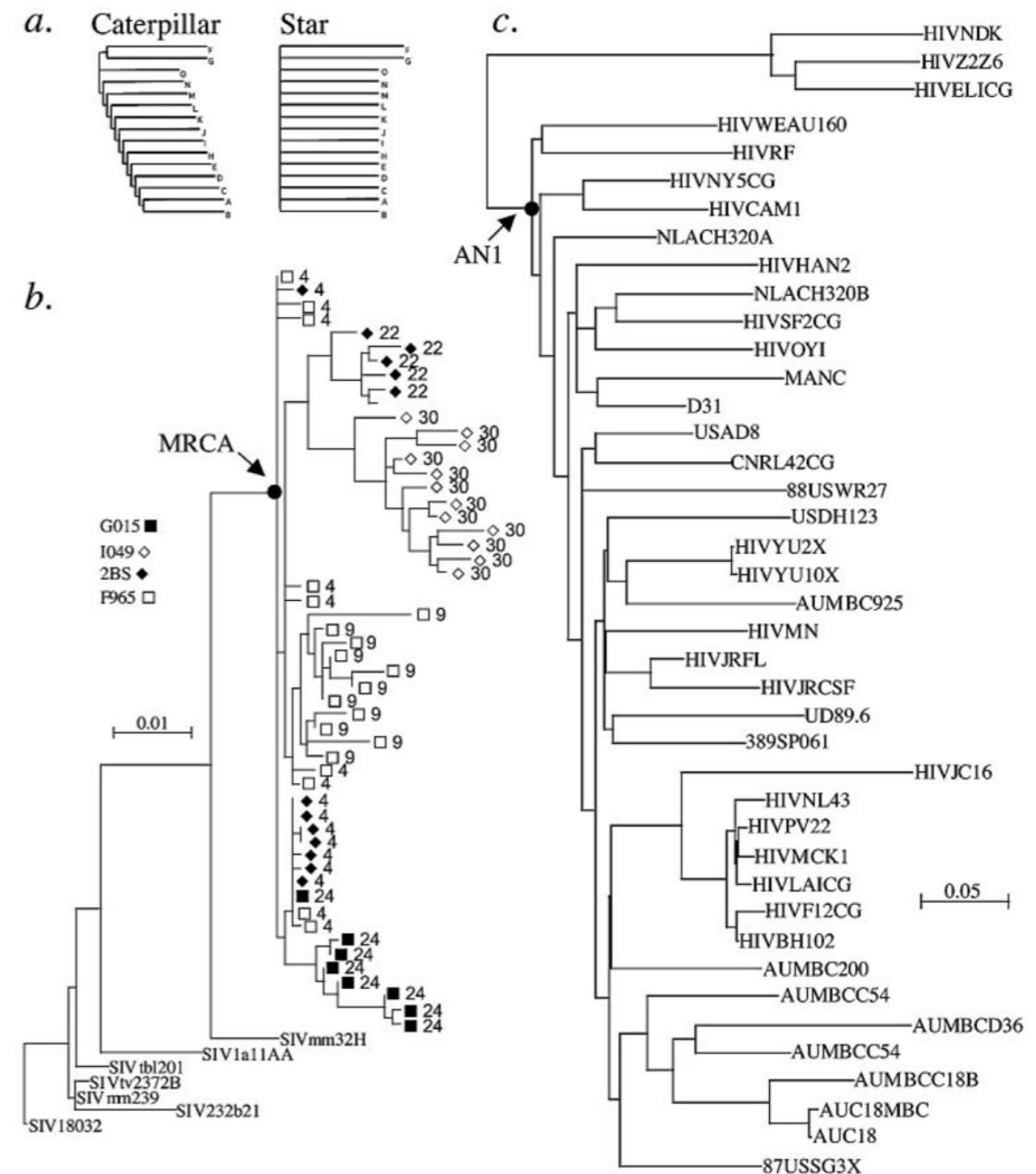
JOURNAL OF VIROLOGY, Sept. 2005, p. 11214–11224  
0022-538X/05/\$08.00+0 doi:10.1128/JVI.79.17.11214–11224.2005  
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

## Human Immunodeficiency Virus Type 1 Subtype B Ancestral Envelope Protein Is Functional and Elicits Neutralizing Antibodies in Rabbits Similar to Those Elicited by a Circulating Subtype B Envelope

N. A. Doria-Rose,<sup>1,2,3‡</sup> G. H. Learn,<sup>2‡§</sup> A. G. Rodrigo,<sup>2‡§</sup> D. C. Nickle,<sup>2</sup> F. Li,<sup>2¶</sup> M. Mahalanabis,<sup>1,2</sup> M. T. Hensel,<sup>3</sup> S. McLaughlin,<sup>2</sup> P. F. Edmonson,<sup>4||</sup> D. Montefiori,<sup>6</sup> S. W. Barnett,<sup>7</sup> N. L. Haigwood,<sup>1,2,3</sup> and J. I. Mullins<sup>2,4,5\*</sup>



Vol. 79, No. 17



# Reconstrucción de estados ancestrales

- ¿Cuál es la secuencia ancestral en el nodo x?

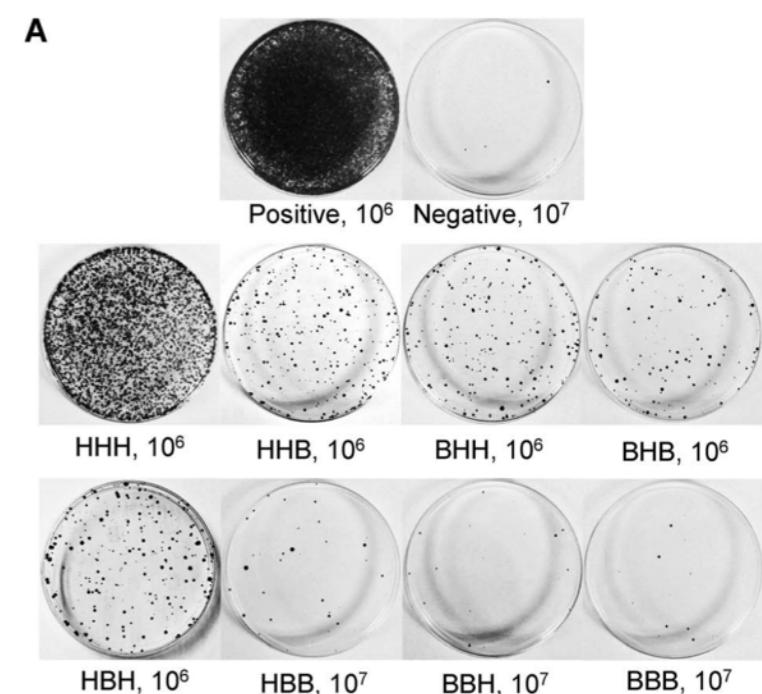
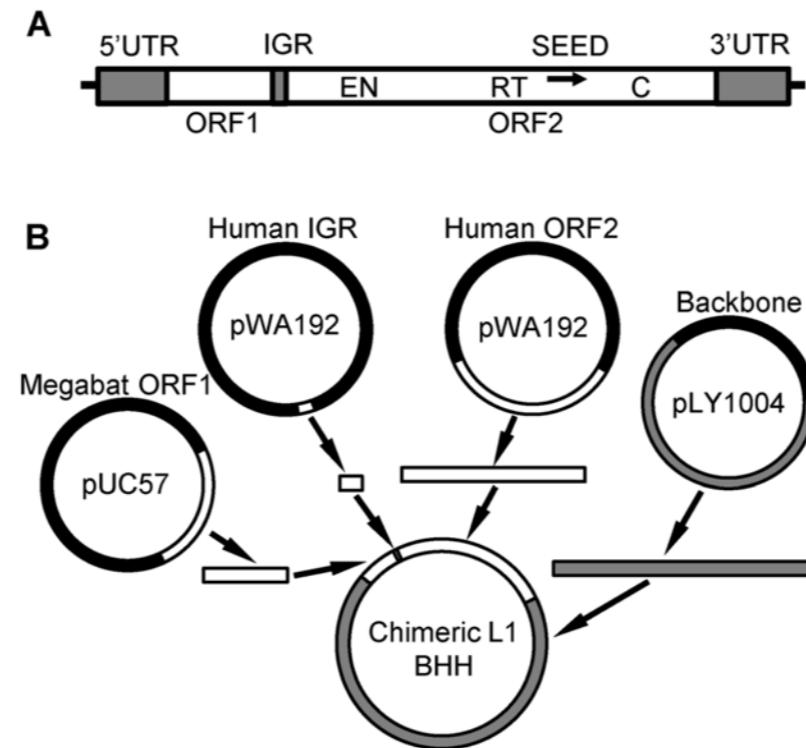
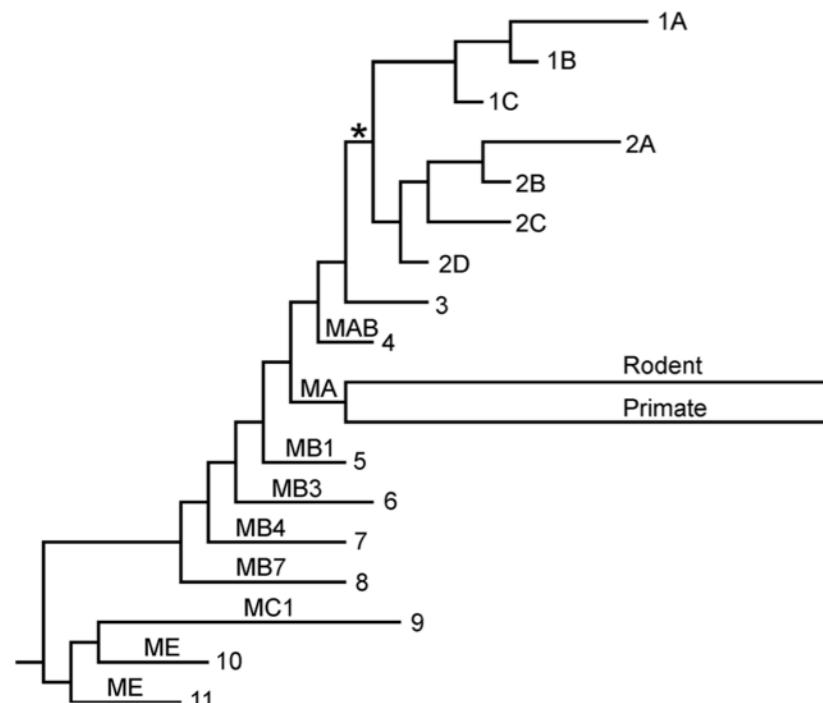
OPEN ACCESS Freely available online

PLOS GENETICS

## Reviving the Dead: History and Reactivation of an Extinct L1

Lei Yang<sup>1,2\*</sup>, John Brunsfeld<sup>1</sup>, LuAnn Scott<sup>1</sup>, Holly Wichman<sup>1,2\*</sup>

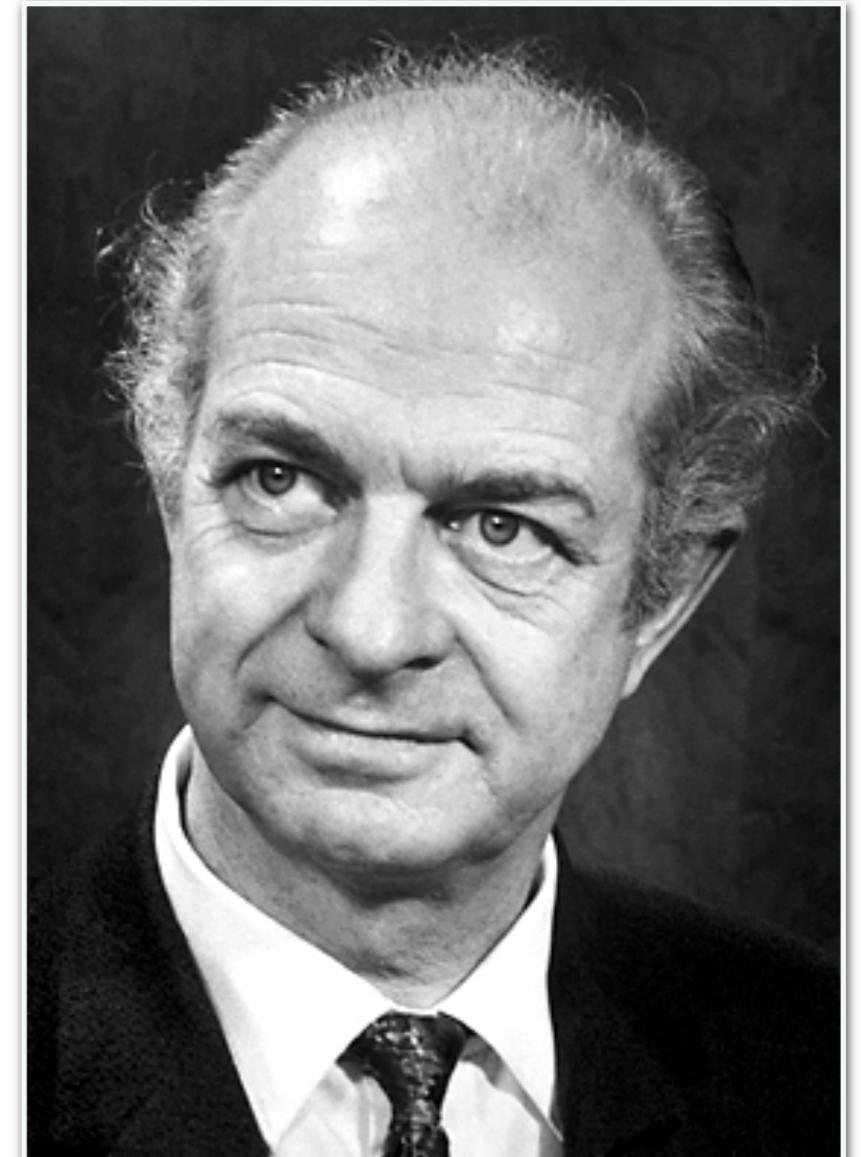
**1** Department of Biological Sciences, University of Idaho, Moscow, Idaho, United States of America, **2** Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, United States of America



# Relojes moleculares

*The best way to have a good idea is to have lots of ideas.*

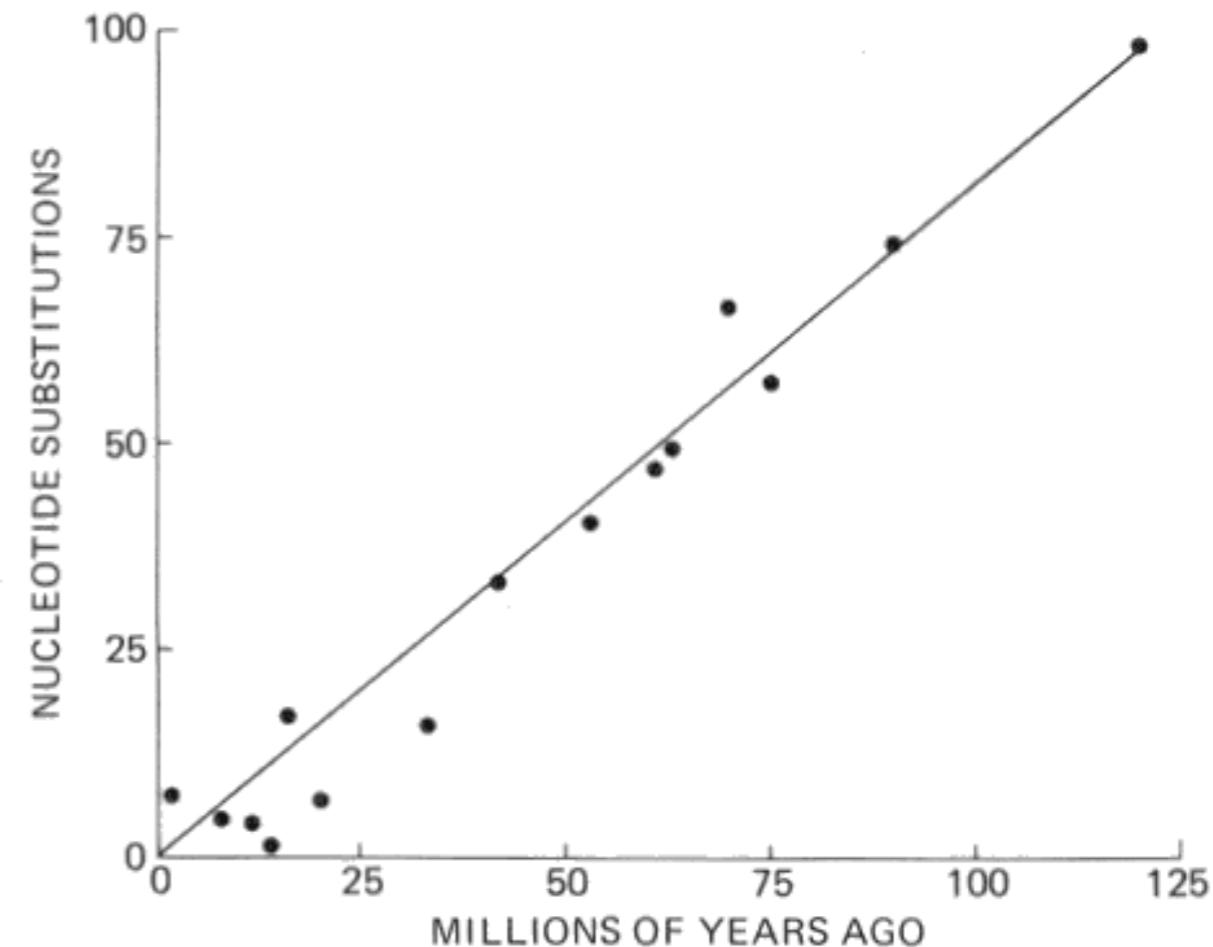
- Emile Zuckerkandl and Linus Pauling
- Basados en la teoría neutral evolutiva
- Tasa de sustitución es constante en el tiempo



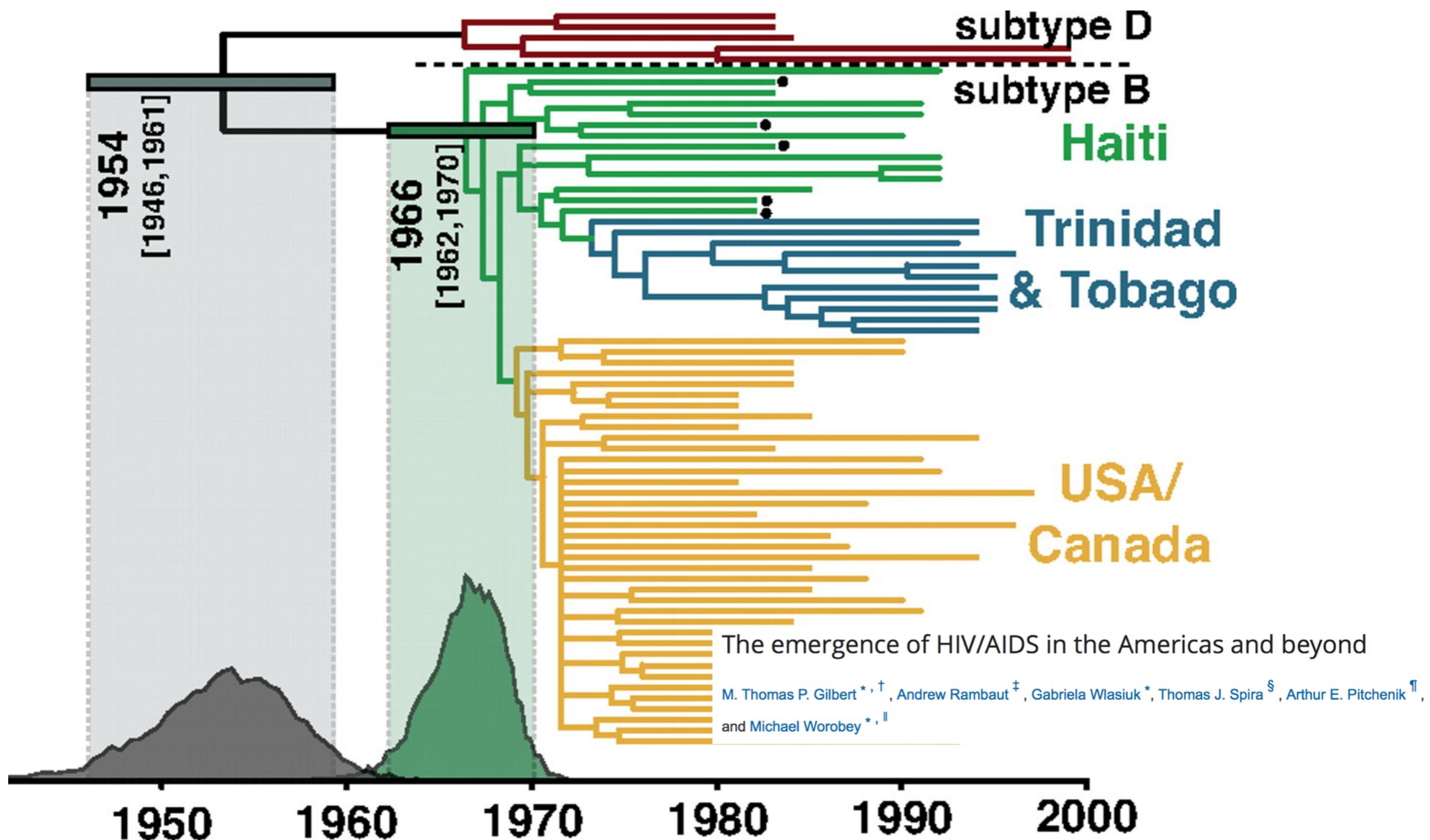
One person, Linus Pauling, has won two undivided Nobel Prizes. In 1954 he won the Prize for Chemistry. Eight years later he was awarded the Peace Prize for his opposition to weapons of mass destruction.

# Relojes moleculares

- Tasa de sustitución es constante en el tiempo
- Se sabe que no es así pero no importa
- Modelos más complejos pueden acomodar variación en tasas de sustitución



# Relojes moleculares



# Relojes moleculares

## Divergencia entre Neandertales y humanos modernos

