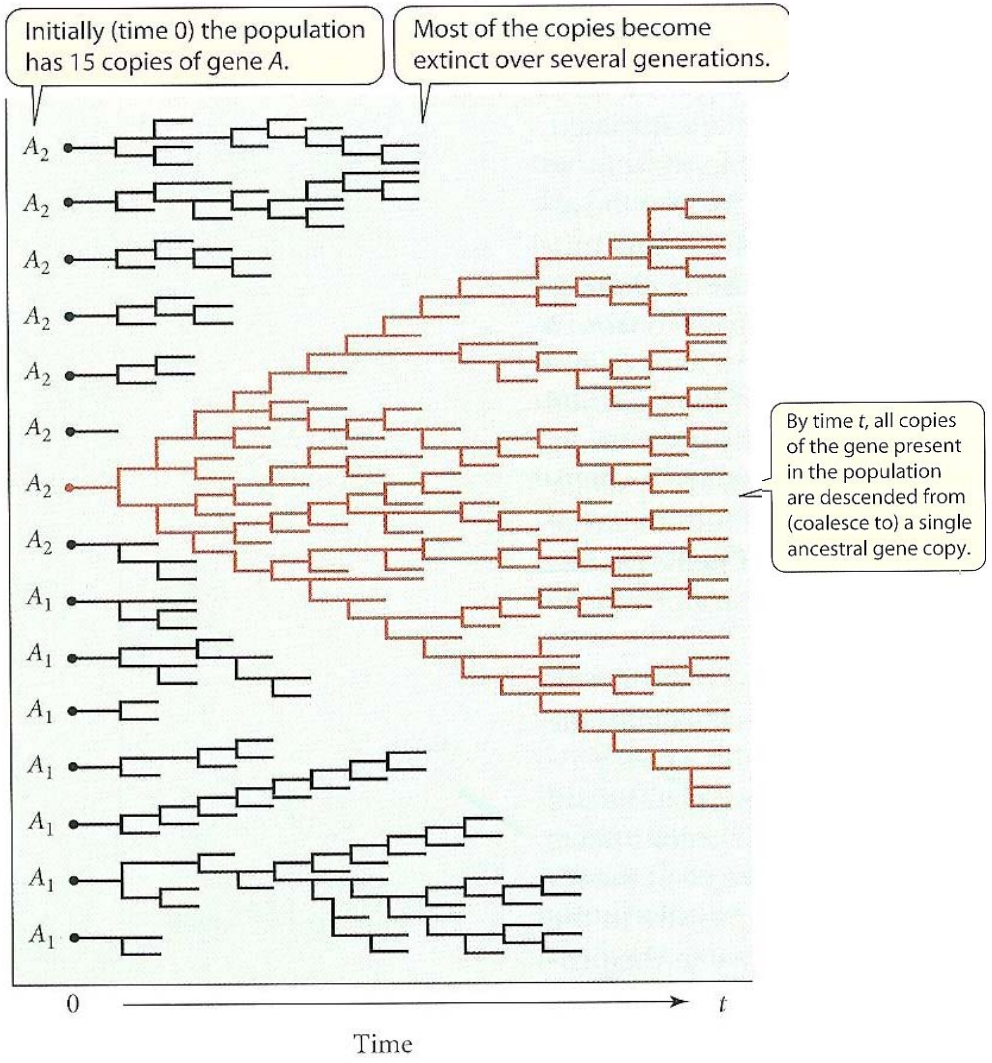
La teoría de coalescencia

- Proceso estocástico
- Es un linaje de alelos proyectado hacia el pasado (su ancestro común mas reciente)
- Aproximación básica usando el modelo de Wright-Fisher
- Modelo probabilístico de una genealogía de una muestra de n genes tomados al azar de una población grande
- Es el avance mas importante de la genética de poblaciones de las últimas 3 décadas. Su uso se refiere a un análisis estadístico ruguroso de diferentes modelos usando datos de poblaciones naturales
- Surgió como una necesidad de estimar parámetros del pasado usando muestras de poblaciones actuales
- Trabajo original de: Kingman, *J Appl Prob* 19A:27, 1982

Deriva génica: visión del pasado al presente (genética de poblaciones retrospectiva)



Una visión desde la replicación del ADN (Templeton)

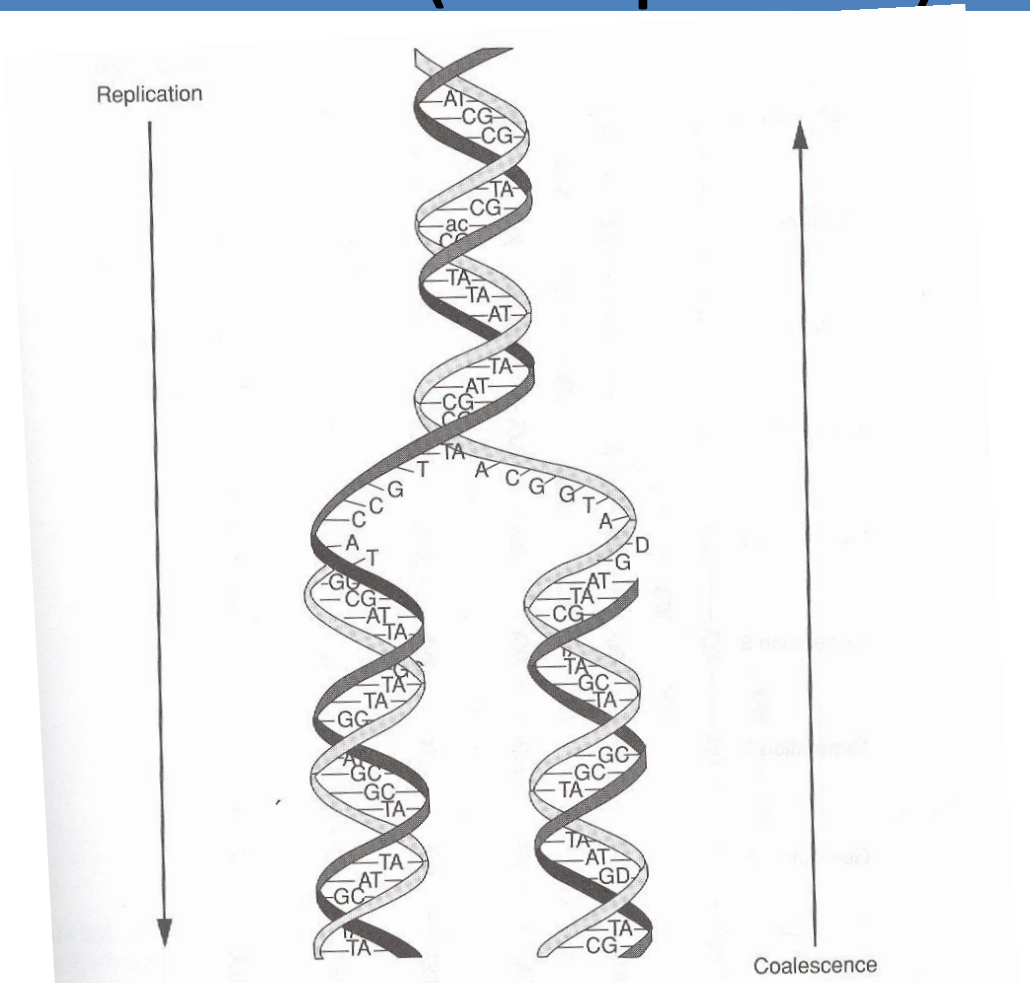


Figure 5.7. DNA replication and its time-inverse coalescence.

Los tiempos de coalescencia son variables aleatorias distribuidos exponencialmente

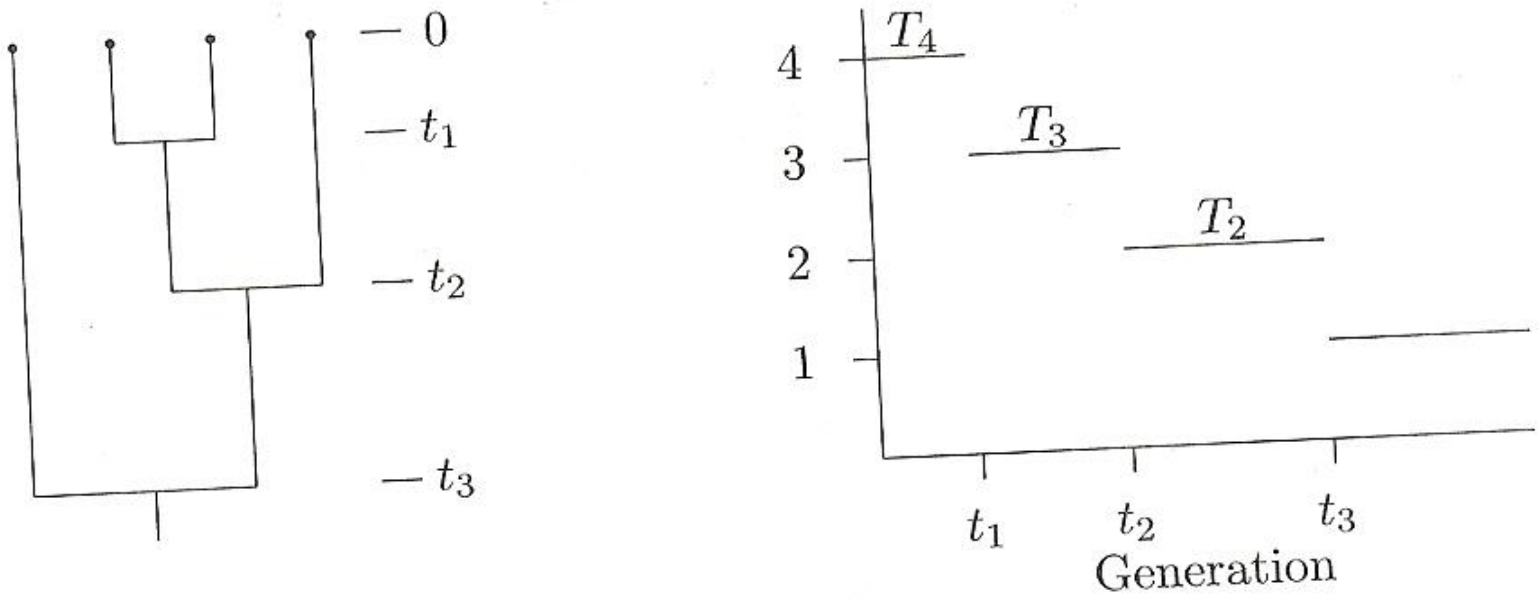
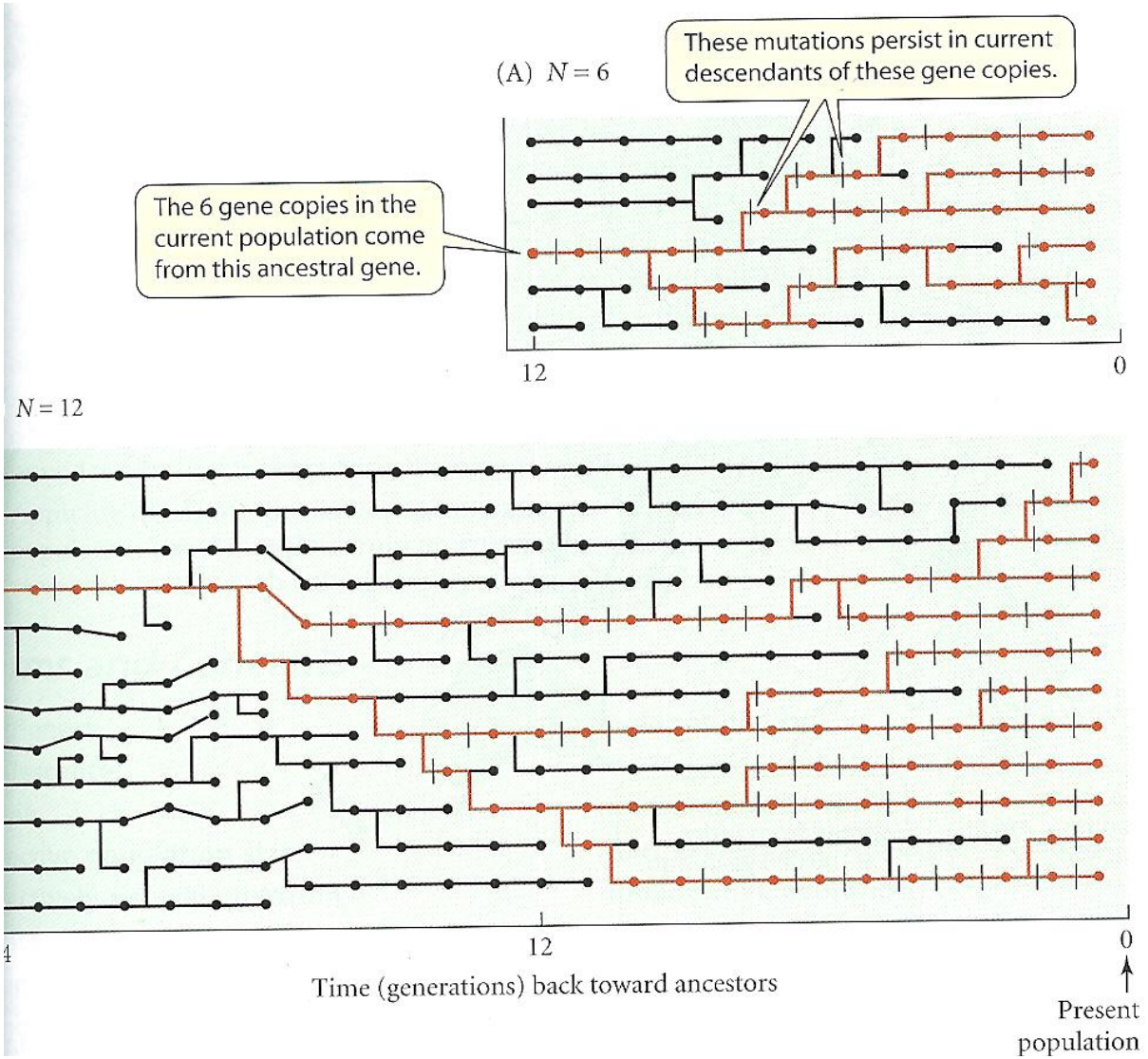
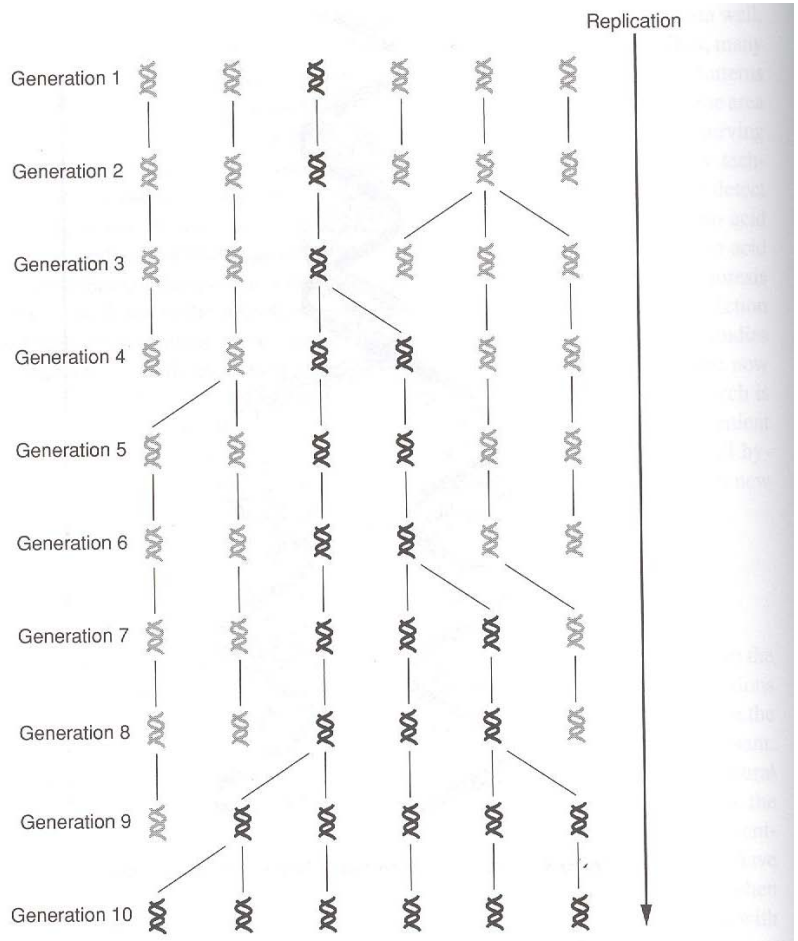


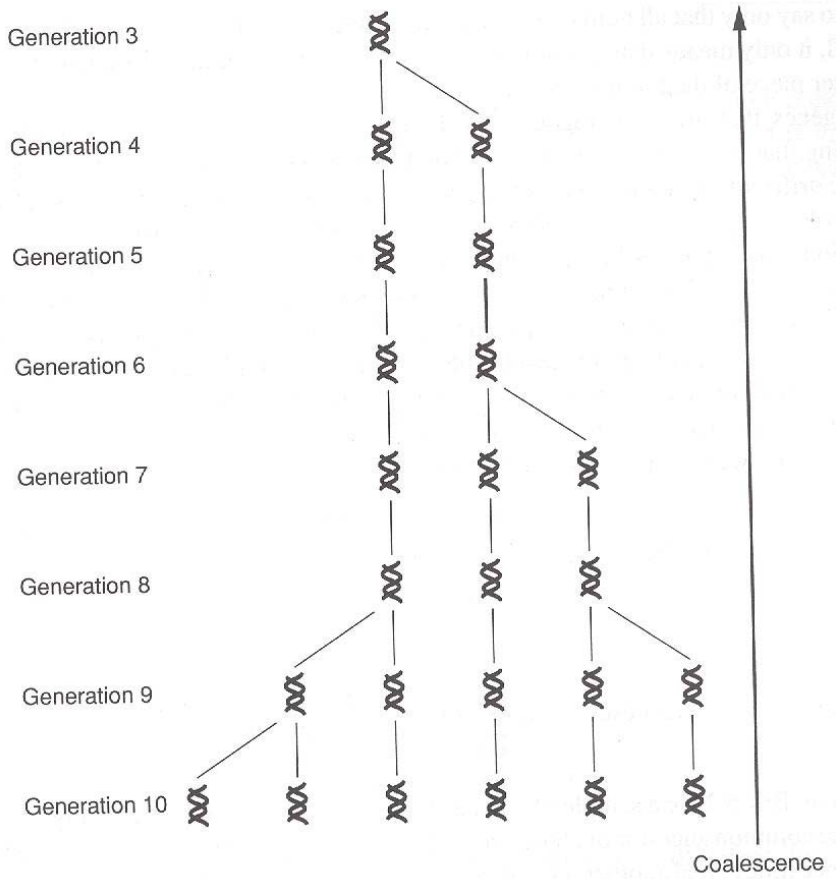
Figure 2.7: On the left is an example of a coalescent for four alleles. On the right is a graph showing the size of the coalescent as a function of time measured backward.

El tiempo de coalescencia (hasta el MRCA) depende de N



Si tomamos 6 haplotipos en 10 generaciones, un resultado podría ser este





Con los haplotipos que no dejaron descendencia (replicación) y sin ellos coalescencia

La probabilidad de que 2 genes coalezcan

P(2 linajes tengan el mismo padre)

y coalezcan	$1/ 2N$
no lo hagan	$1- 1/ 2N$

Si consideramos un tercer linaje, la probabilidad de que siendo desciendan

La probabilidad de que 3 genes coalezcan

P(2 linajes tengan el mismo padre)

Y coalezcan	$1/ 2N$
No lo hagan	$1- 1/ 2N$

Si consideramos un tercer linaje, la probabilidad de que desciendan de diferentes padres sería (1-1/2N) x ((2N-2)/2N) o (1-1/2N) x (1-2/2N)

En general para n linajes

Probabilidad de que n linajes tengan
n padres diferentes en la generación previa sería:

$$\Pr(n) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{2N}\right) \qquad (8.11a)$$

Y la probabilidad de que los n linajes o genes muestreados tengan n ancestros
t generaciones atrás sería:
$[\Pr(n)]^t$

Probabilidad de que dos genes o linajes no coalezcan en *t* generaciones y coalezcan en la generación *t*+1

$$\Pr(2)^t[1 - \Pr(2)] = \left(1 - \frac{1}{2N}\right)^t \frac{1}{2N}$$

Esta ecuación se puede expresar también como:

$$\text{Prob(coalescence at } t) = \left(1 - \frac{1}{xN_{ef}}\right)^{t-1} \left(\frac{1}{xN_{ef}}\right)$$

Donde *x* es la ploidía y N_{ef} esel tamaño efectivo de la población

El tiempo promedio a la coalescencia sería entonces

$$\text{Average time to coalescence} = \sum_{t=1}^{\infty} t \left(1 - \frac{1}{xN_{ef}}\right)^{t-1} \left(\frac{1}{xN_{ef}}\right) = xN_{ef}$$

Donde x depende de la ploidía y la forma de herencia
y N_{ef} el tamaño efectivo de la población

Probabilidad de que *n* genes o linajes tengan *n*-1 ancestros, *t*+1 generaciones atrás

$$\Pr(n)^t[1-\Pr(n)] = \left[\prod_{i=1}^{n-1}(1-\frac{i}{2N})\right]^t \frac{1}{2N} \qquad (8.11b)$$

El n-colescente

$$\text{Number of pairs of genes} = \binom{n}{2} = \frac{n!}{(n-2)!2!} = \frac{n(n-1)}{2}$$

$$\text{Prob(coalescence in previous generation)} = \binom{n}{2}\frac{1}{xN} = \frac{n(n-1)}{2xN}$$

Hence,

$$\text{Prob(no coalescence in previous generation)} = 1 - \frac{n(n-1)}{2xN}$$

$$\text{Prob(first coalescence in } t \text{ generations)} = \left(1 - \frac{n(n-1)}{2xN}\right)^{t-1}\frac{n(n-1)}{2xN}$$

and the expected time to the first coalescence is

$$E(\text{time to first coalescence}) = \sum_{t=1}^{\infty} t\left(1 - \frac{n(n-1)}{2xN}\right)^{t-1}\frac{n(n-1)}{2xN} = \frac{2xN}{n(n-1)}$$

Tiempo esperado durante el que hay *n* linajes diferentes

$$E(T_n) = \frac{4N}{n(n-1)}$$

La variación entre muestras usando los mismos parámetros en la simulación

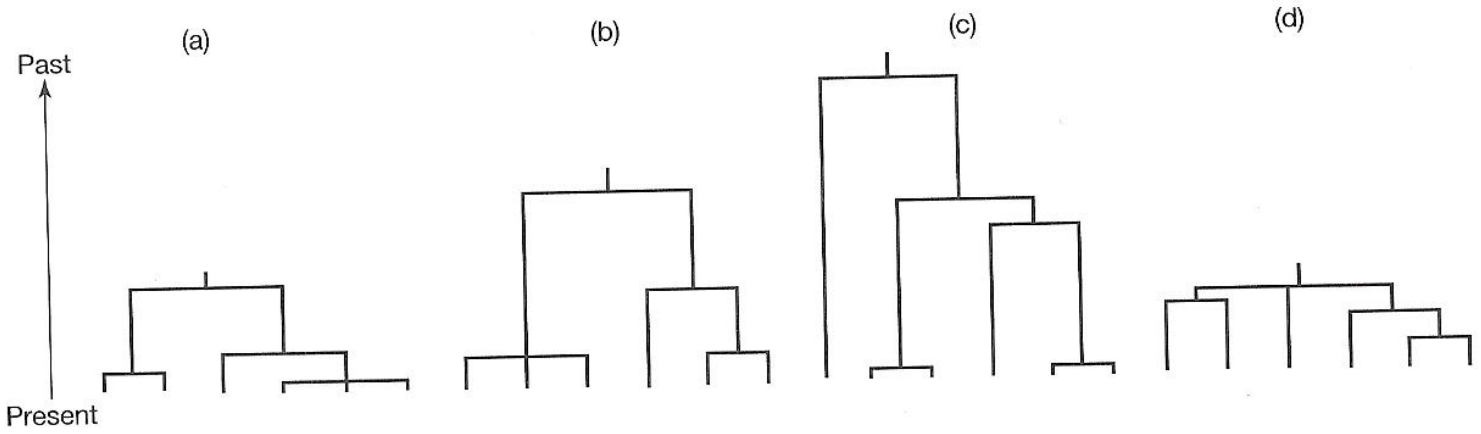


Figure 8.14. Four examples of coalescent trees for a sample size of $n = 6$ drawn on the same scale (Nordborg, 2001). Labels (1 to 6) indicating sample number should be randomly assigned to the tips of the tree.

Tiempo esperado a la coalescencia de los *n* linajes o genes

$$E(t) = \sum_{i=2}^{n} E(T_i)$$

$$= 4N\left(1 - \frac{1}{n}\right)$$

$$E(\text{time to coalescence of all } n \text{ genes}) = \sum_{k=1}^{n-1} \frac{2xN}{(n-k+1)(n-k)} = 2xN\left(1 - \frac{1}{n}\right)$$

Para 5 genes el tiempo esperado es 3.2N generaciones

El tiempo total de la coalescencia (T_c en Gillespie) sería:

$$E\{T_c\} = \sum_{i=2}^{n} iE\{T_i\} = 4N \sum_{i=2}^{n} \frac{1}{i-1}.$$

Partiendo de:

$$T_c = \sum_{i=2}^{n} iT_i,$$ y de:

$$E\{T_i\} = \frac{4N}{i(i-1)},$$

Para generar estas genealogías podemos usar números aleatorios de una distribución uniforme entre 0 y 1 y:

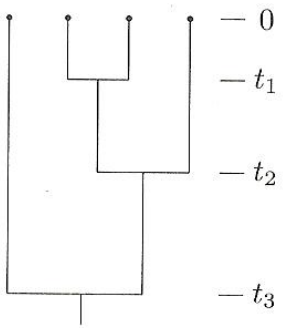
$$T_i = \frac{-2\ln(1-x)}{i(i-1)}$$

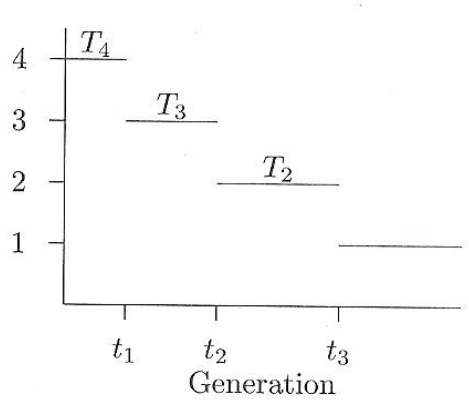
Por ejemplo si tenemos una muestra de 6 linajes y el primer número aleatorio es x = 0.22, entonces T_6 = 0.0166, si el segundo es x = 0.57, T_5 = 0.0844, etc.

Cada simulación genera una topología diferente para 6.

Los tiempos de coalescencia y el número de mutaciones en la muestra

El tiempo total de coalescencia





Para 4 linajes

El número de mutaciones

$$uE\{T_c\} = \theta(11/6)$$

El número de sitios segregantes esperado para 4 linajes sería

Si introducimos mutación…1

Caso 1. Coalescencia primero, mutación después

$$\begin{aligned}
\text{Prob(coalescence before mutation)} &= \text{Prob(identity by descent)} \\
&= \left(1 - \frac{1}{xN_{ef}}\right)^{t-1} \left(\frac{1}{xN_{ef}}\right) (1-\mu)^{2t} \\
&= \text{Prob(no coalescence for } t-1 \text{ generations)} \\
&\quad \times \text{Prob(coalescence at generation } t\text{)} \\
&\quad \times \text{Prob(no mutation in } 2t \text{ DNA replications)}
\end{aligned}$$

Si introducimos mutación…2

Caso 2. Mutación primero, coalescencia después

$$\text{Prob(mutation before coalescence)} = \left(1 - \frac{1}{xN_{ef}}\right)^{t} 2\mu(1-\mu)^{2t-1}$$

Si introducimos mutación…3

Prob(mutation before coalescence | mutation or coalescence)

$$= \frac{2\mu(1-\mu)^{2t-1}\left(1-1/xN_{ef}\right)^{t}}{2\mu(1-\mu)^{2t-1}\left(1-1/xN_{ef}\right)^{t}+\left[1/\left(xN_{ef}\right)\right](1-\mu)^{2t}\left(1-1/xN_{ef}\right)^{t-1}}$$

$$= \frac{2xN_{ef}\mu-2\mu}{2xN_{ef}\mu-3\mu+1} \qquad (5.13)$$

If $\mu \ll N_{ef}\mu$ (i.e., a large inbreeding effective size) and defining $\theta = 2xN_{ef}$, equation 5.13 simplifies to

Prob(mutation before coalescence | mutation or coalescence)

$$= \frac{2xN_{ef}\mu-2\mu}{2xN_{ef}\mu-3\mu+1} \approx \frac{2xN_{ef}\mu}{2xN_{ef}\mu+1} = \frac{\theta}{\theta+1} \qquad (5.14)$$

Esta última ecuación es igual a la ecuación de la heterocigosidad en el Equilibrio que derivamos antes y donde theta = $2xN_{ef}$

Si introducimos mutación…4

$$E\{S_n\} = uE\{T_c\} = \theta \sum_{i=2}^{n} \frac{1}{(i-1)},$$

Donde….

$$E\{T_c\} = \sum_{i=2}^{n} iE\{T_i\} = 4N \sum_{i=2}^{n} \frac{1}{i-1}.$$

Estimación del MRCA para algunos genes en humanos

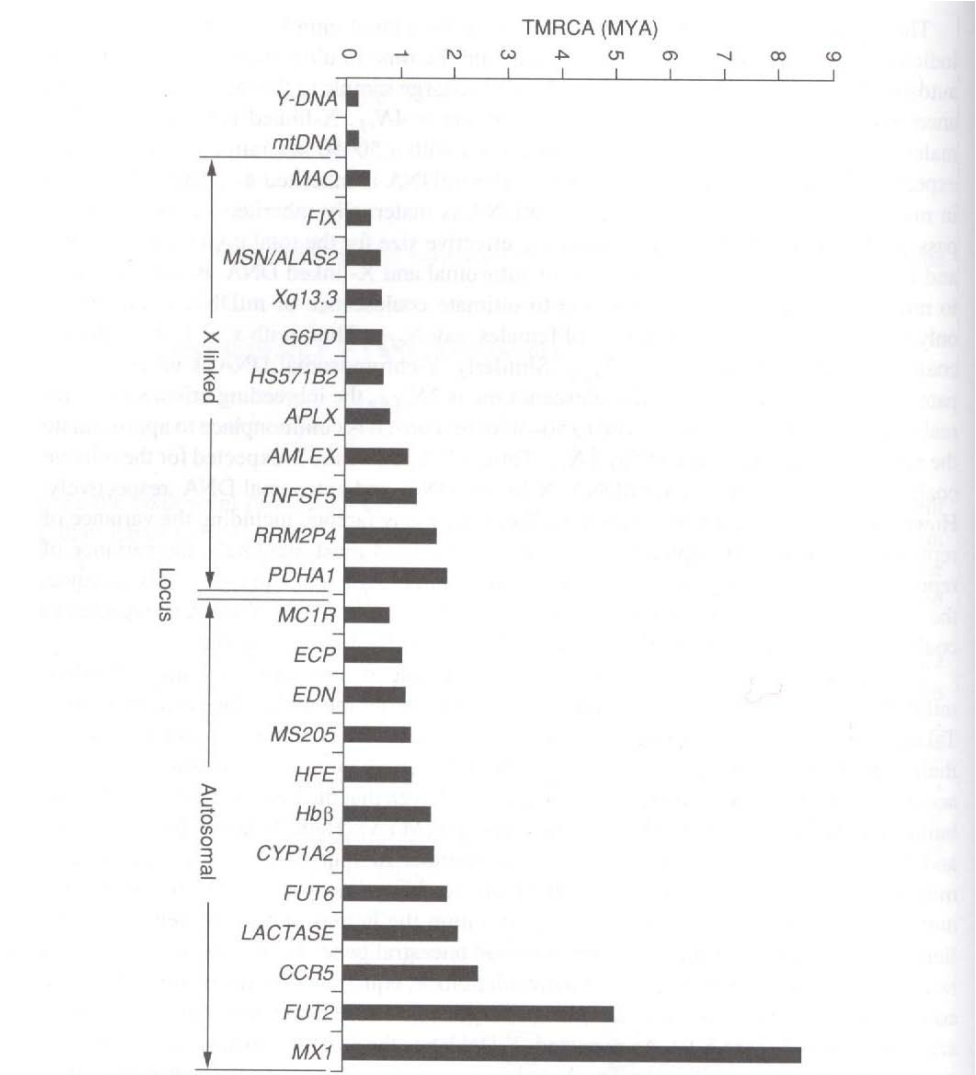


Figure 5.10. Estimated coalescent times (time to the most recent common ancestor, or TMRCA) for 25 human DNA regions. Details and references for the DNA regions studied are given in Templeton (2005).

Y…

$$\hat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{3} \cdots + \frac{1}{n-1}}$$

Sería un buen estimador de:

$$\theta = 4Nu.$$

Muestras usando el modelo de Wright-Fisher

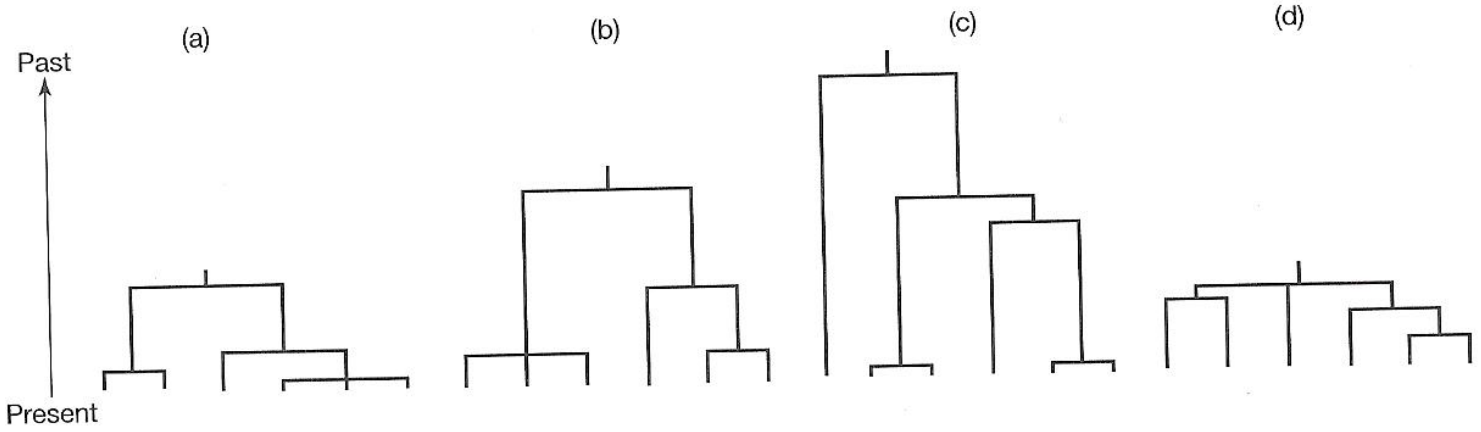


Figure 8.14. Four examples of coalescent trees for a sample size of $n = 6$ drawn on the same scale (Nordborg, 2001). Labels (1 to 6) indicating sample number should be randomly assigned to the tips of the tree.

La topologíade la coalescencia nos ayuda a entender la historia de la muestra demográfica y evolutivamente si tienen una baja probabilidad de ser neutras

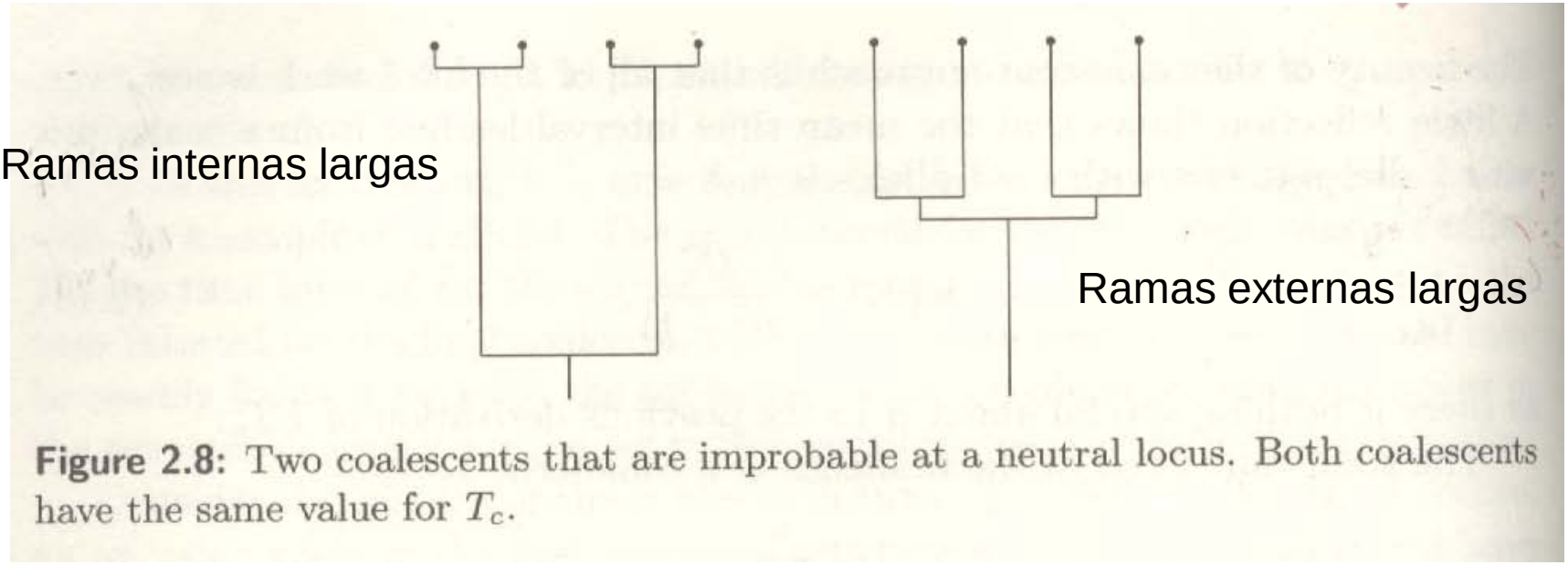


Figure 2.8: Two coalescents that are improbable at a neutral locus. Both coalescents have the same value for T_c.

Si hay 4 mutaciones, en el caso de la izquierda habría dos haplotipos mientras que en el caso de la derecha habrá 4 haplotipos

En el caso neutro, con mutación y deriva, se espera…

$$\pi = E\left\{\sum_{i=1}^{\infty} 2p_i(1-p_i)\right\} = \theta.$$

Los estimados de π y θ deben de ser iguales en un modelo Wright-Fisher con mutación

$$\hat{\pi} = \frac{n}{n-1} \sum_{i=1}^{S_n} 2\hat{p}_i(1-\hat{p}_i),$$

$$\hat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{3} \cdots + \frac{1}{n-1}}$$

Prueba de Tajima

Hipótesis bajo un modelo Wright-Fisher con mutación

$$E\{\hat{\pi}\} = \theta.$$

Estadístico de Tajima

$$D_T = \frac{\hat{\pi} - \hat{\theta}}{C},$$

Donde,

$$a_1 = \sum_{i=1}^{n-1} i^{-1}, \; a_2 = \sum_{i=1}^{n-1} i^{-2}$$

$$b_1 = (n+1)/[3(n-1)]$$

$$b_2 = 2(n^2+n+3)/[9n(n-1)]$$

$$c_1 = b1 - (1/a_1)$$

$$c_2 = b_2 - (n+2)/(a_1 n) + a_2/a_1^2.$$

With these in hand,

$$C = \sqrt{(c_1/a_1)S_n + [c_2/(a_1^2+a_2)]S_n(S_n-1)}.$$

Frecuencias de los sitios polimórficos

- Definamos ξ_i como el número de sitios segregantes donde la base mutante está en i secuencias y la ancestral en n-i.

- Por otro lado η_1 es el número de sitios segregantes donde la frecuencia del alelo es i.

Callitropsis guadalupensis, polimorfismos en dos regiones del cloroplasto

	*trn*S-*trn*G										*trn*L-*trn*F		
Posición / Haplotipo	105	113	301	333	381-410	411 - 414	442	510	746	842	1273	1316	138 - 139
H01	C	C	C	T	TATATATATATATATATATATATATATATA	-- --	T	G	T	T	A	C	***
H02	.	A	.	.	TATATATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H03	.	A	.	.	TATATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H04	.	A	.	.	TATATATATA - ATATATATATATA	-- --	.	.	.	.	.	.	.
H05	.	A	.	.	TATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H06	.	A	.	.	TATATATATATATATATATATA	-- --	.	.	.	C	.	.	.
H07	.	A	A	C	TATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H08	.	A	.	C	TATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H09	.	.	.	.	TATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H10	.	.	.	.	TATATATATATATATATATATA	-- --	C	.	.	.	.	.	.
H11	.	A	.	.	TATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H12	.	.	.	.	TATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H13	.	.	.	.	TATATATAT - TATATATATA	-- --	.	.	.	.	.	.	.
H14	.	.	.	.	TATATATATATAT - TATATA	-- --	.	.	.	.	.	.	.
H15	.	A	.	.	TATATATATATATATATA	-- --	.	.	.	.	.	.	.
H16	.	.	.	.	TATATATATATATATATA	-- --	.	.	.	.	.	.	.
H17	.	.	.	.	TATATATATATATATATA	-- --	C	.	.	.	.	T	.
H18	.	.	.	.	TATATATATATATATATA	TT	.	.	.	.	.	.	.
H19	.	A	.	.	TATATATATATATATA	-- --	.	.	.	.	.	.	.
H20	.	.	.	.	TATATATATATATATA	-- --	.	.	.	.	.	.	---
H21		.	.	.	TATATATATATA	-- --	.	.	.	.	.	.	.
C. sargentii	T	.	.	.	TATATATATATATA	TTTT	.	T	G	.	G	.	.
Juniperus	.	.	.	.	-- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --	-- --	.	.	G	.	G	.	.

Datos de Rosas, 2008

Distribución de frecuencias de los sitios nucleotídicos polimórficos

Singletons = 105, 301, 842, 1316
Doubletons = 333, 442
Tenton = 113

	*trn*S-*trn*G										*trn*L-*trn*F		
Posición / Haplotipo	105	113	301	333	381-410	411 - 414	442	510	746	842	1273	1316	138 - 139
H01	C	C	C	T	TATATATATATATATATATATATATATATA	-- --	T	G	T	T	A	C	***
H02	.	A	.	.	TATATATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H03	.	A	.	.	TATATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H04	.	A	.	.	TATATATATA - ATATATATATATA	-- --	.	.	.	.	.	.	.
H05	.	A	.	.	TATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H06	.	A	.	.	TATATATATATATATATATATATA	-- --	.	.	.	C	.	.	.
H07	.	A	A	C	TATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H08	.	A	.	C	TATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H09	.	.	.	.	TATATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H10	.	.	.	.	TATATATATATATATATATATATA	-- --	C	.	.	.	.	.	.
H11	.	A	.	.	TATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H12	.	.	.	.	TATATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H13	.	.	.	.	TATATATAT - TATATATATA	-- --	.	.	.	.	.	.	.
H14	.	.	.	.	TATATATATATAT - TATATA	-- --	.	.	.	.	.	.	.
H15	.	A	.	.	TATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H16	.	.	.	.	TATATATATATATATATATA	-- --	.	.	.	.	.	.	.
H17	.	.	.	.	TATATATATATATATATATA	-- --	C	.	.	.	.	T	.
H18	.	.	.	.	TATATATATATATATATATA	TT	.	.	.	.	.	.	.
H19	.	A	.	.	TATATATATATATATATA	-- --	.	.	.	.	.	.	.
H20	.	.	.	.	TATATATATATATATATA	-- --	.	.	.	.	.	.	---
H21		.	.	.	TATATATATATATA	-- --	.	.	.	.	.	.	.
C. sargentii	T	.	.	.	TATATATATATATA	TTTT	.	T	G	.	G	.	.
Juniperus	.	.	.	.	-- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --	-- --	.	.	G	.	G	.	.

Alelos ancestrales: 113 = C; 301= C; 333 = T, etc.

Prueba de Fu y Li

$$D^* = \frac{S/a_1 - \frac{n-1}{n}\eta_1}{\sqrt{\widehat{\mathrm{Var}}\left[S/a_1 - \frac{n-1}{n}\eta_1\right]}},$$

$$F^* = \frac{\pi - \frac{n-1}{n}\eta_1}{\sqrt{\widehat{\mathrm{Var}}\left[\pi - \frac{n-1}{n}\eta_1\right]}},$$

Donde, η_1 es el número de singletons de la muestra

y, $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$

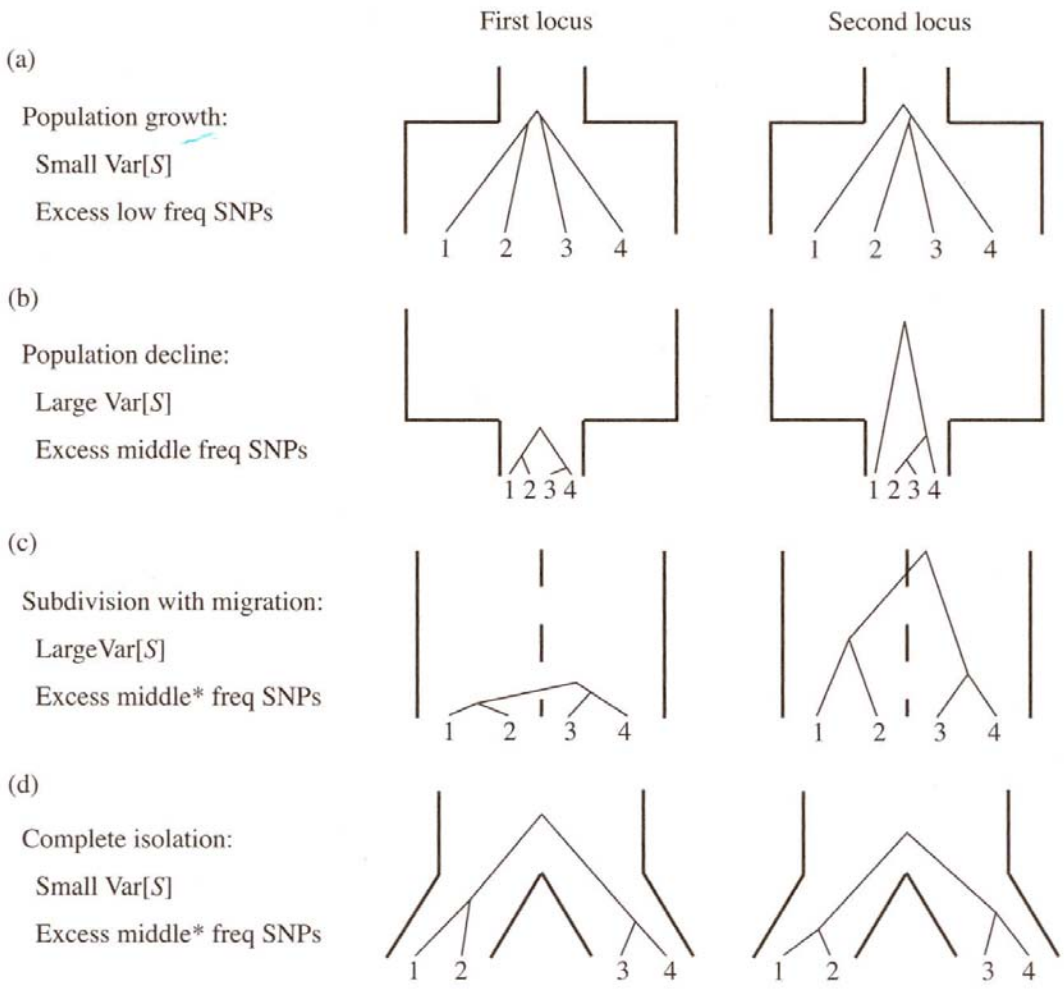
Las relaciones entre los estadísticos de la prueba de Tajima y los de la prueba de Fu y Li

$$S = \sum_{i=1}^{n-1} \xi_i,$$

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i,$$

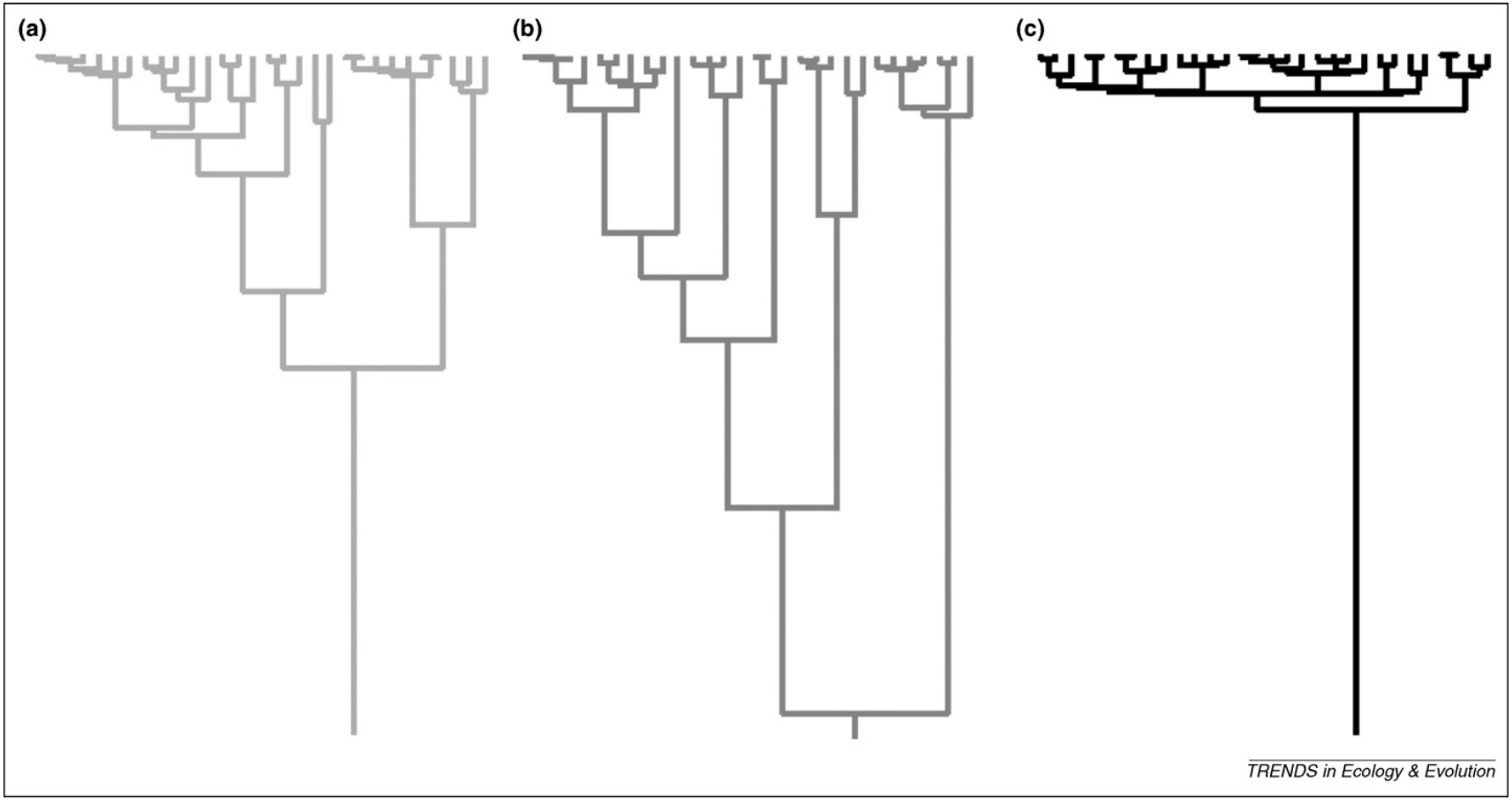
$$\eta_1 = \frac{\xi_1 + \xi_{n-1}}{1 + \delta_{1,n-1}},$$

Efecto de distintas historias demográficas neutrales en los patrones de polimorfismo de ADN



Distintas historias demográficas producen diferentes genealogías

(a) Tamaño constante, (b) Disminución exponencial, (c) Aumento exponencial



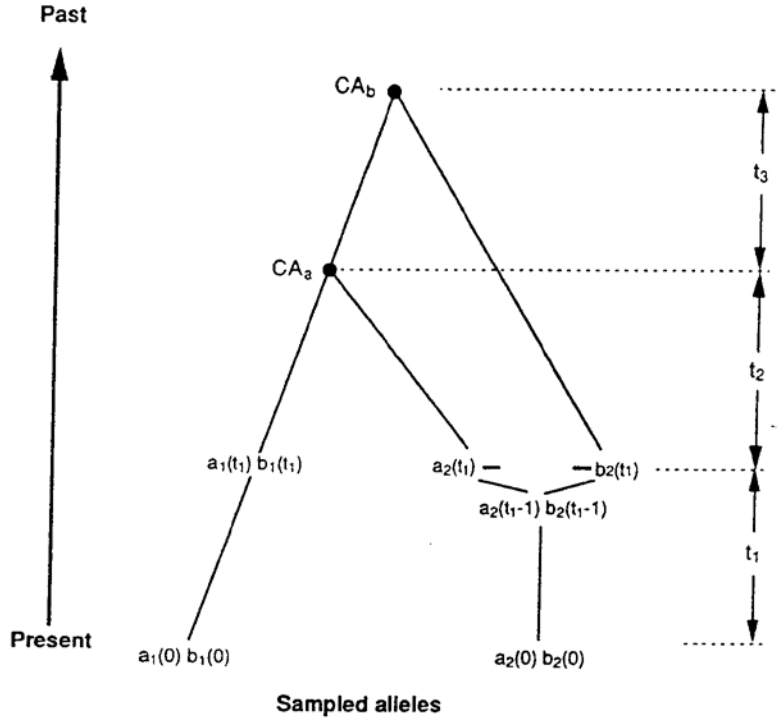
Kuhner, 2009, TREE

Efecto de otros mecanismos evolutivos en la topología de la coalescencia para diferentes regiones del genoma

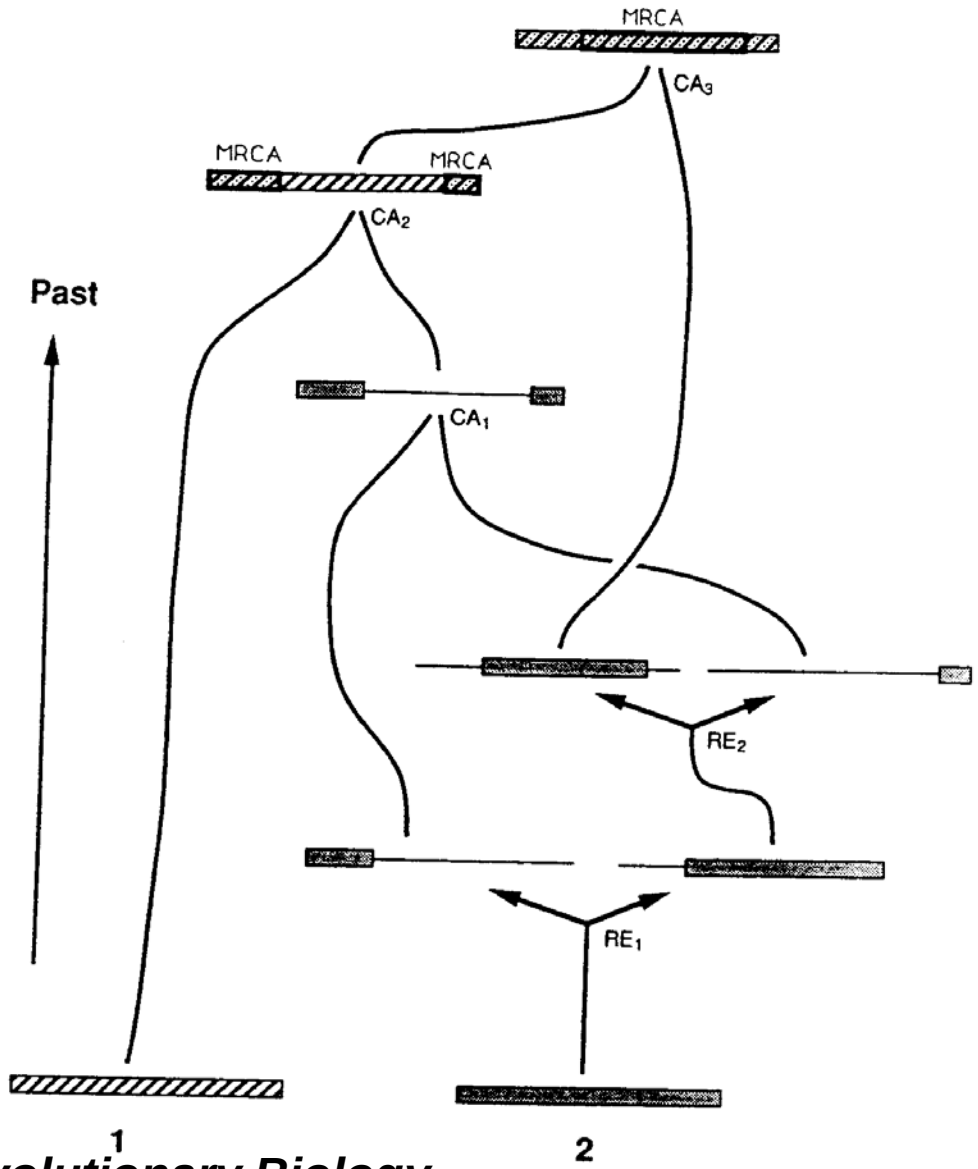
	Whole genome effect	Local effect
Long external branches (Tajima's $D < 0$)	Population growth Very severe bottleneck	Directional selection
Long internal branches (Tajima's $D > 0$)	Population subdivision Less severe bottleneck	Balancing selection Recent population mixing

Coalescencia y recombinación

- La recombinación genera un contexto téorico de la coalescencia que es mas difícil de incorporar en los análisis y generar predicciones.
- En este caso se reconstruyen ARCs (Ancestral Recombination Graphs).

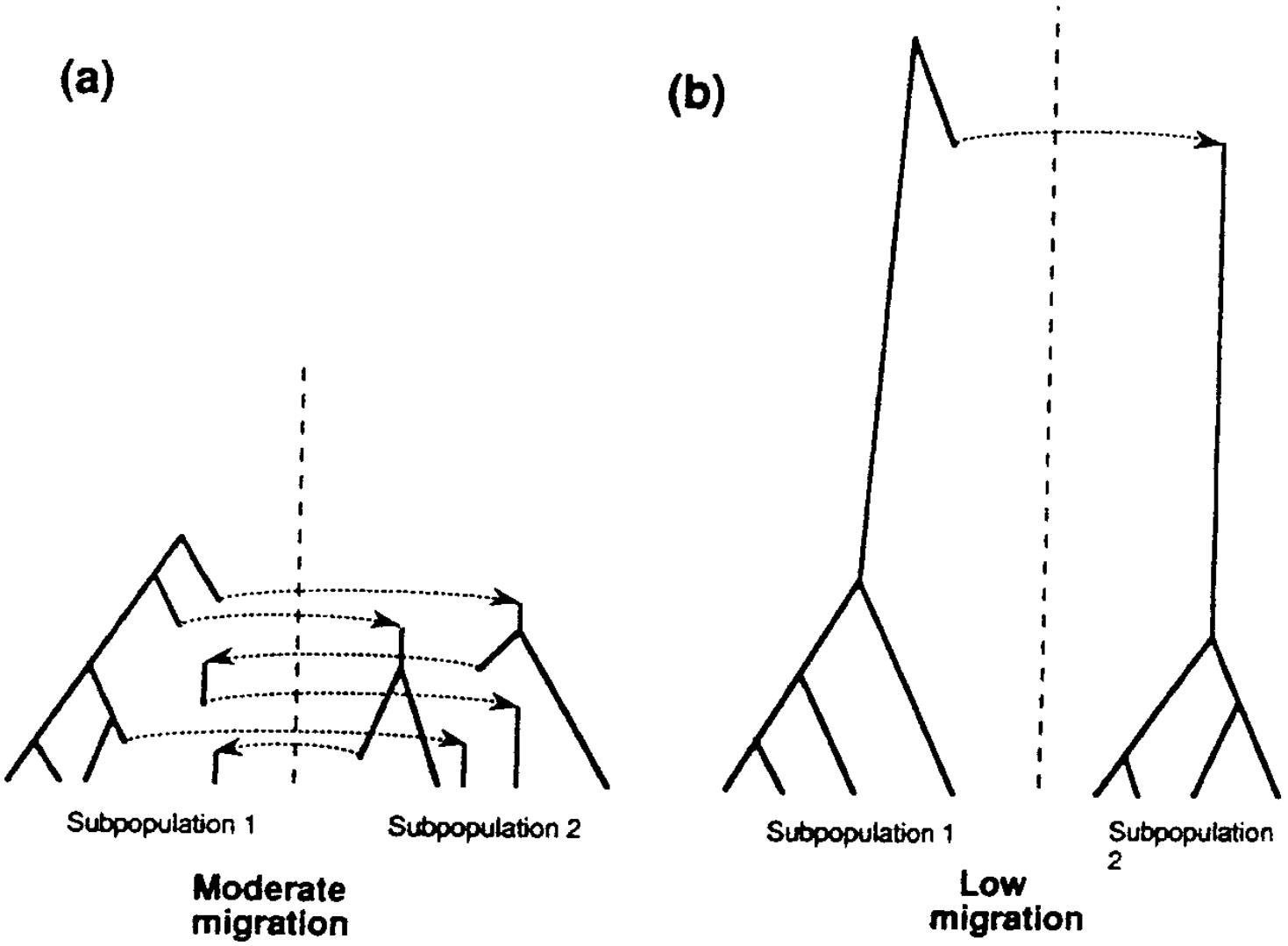


Coalescencia y recombinación…2



Hudson, 1991
Oxford Surveys in Evolutionary Biology

Coalescencia y estructura poblacional: la migración genera coalescencias mas recientes y distancias pareadas mayores



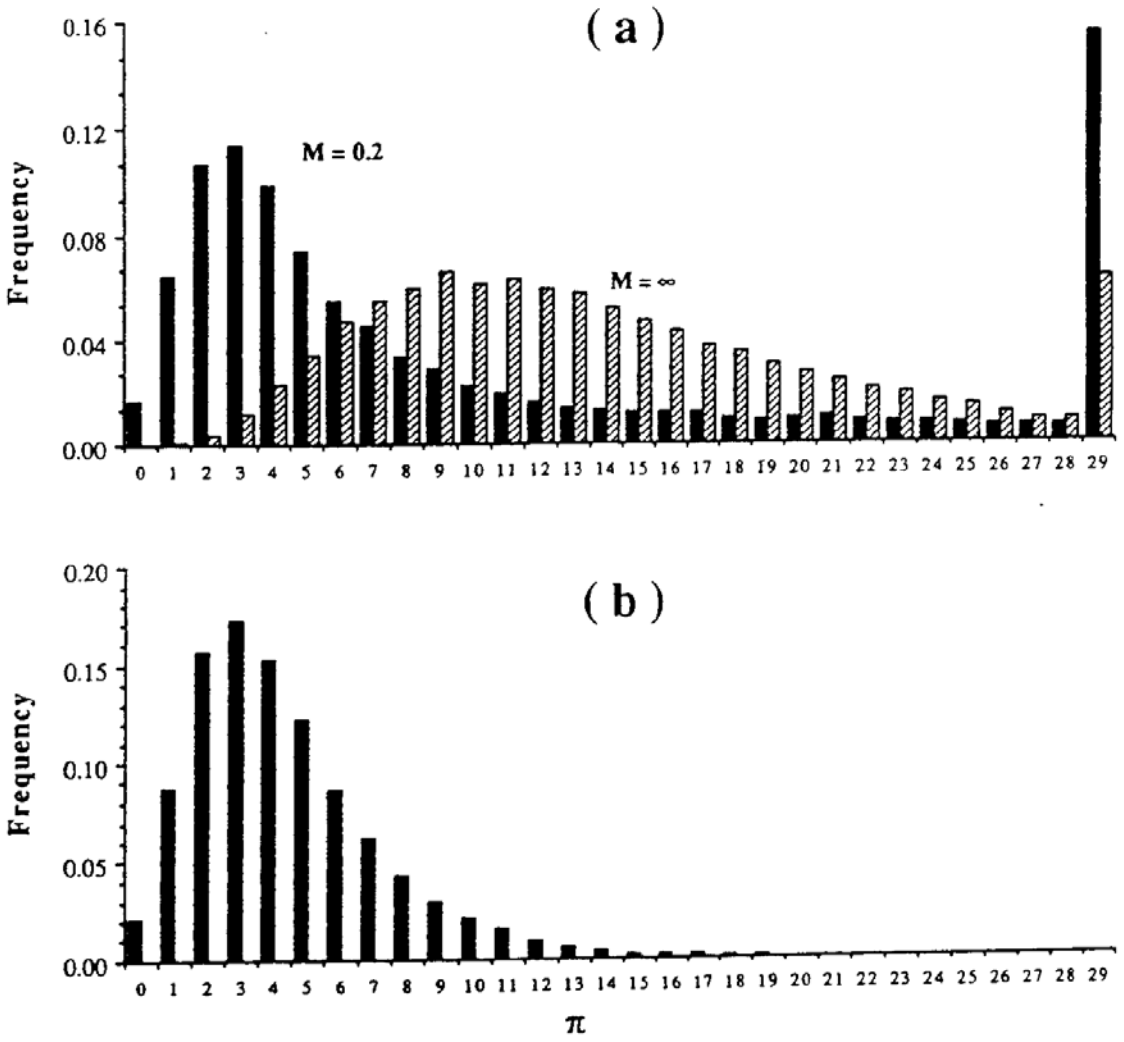
Coalescencia y estructura poblacional: la migración genera coalescencias mas recientes y distancias pareadas mayores…2

(a) Dos poblaciones aisladas

(b) Una población

En ambos casos

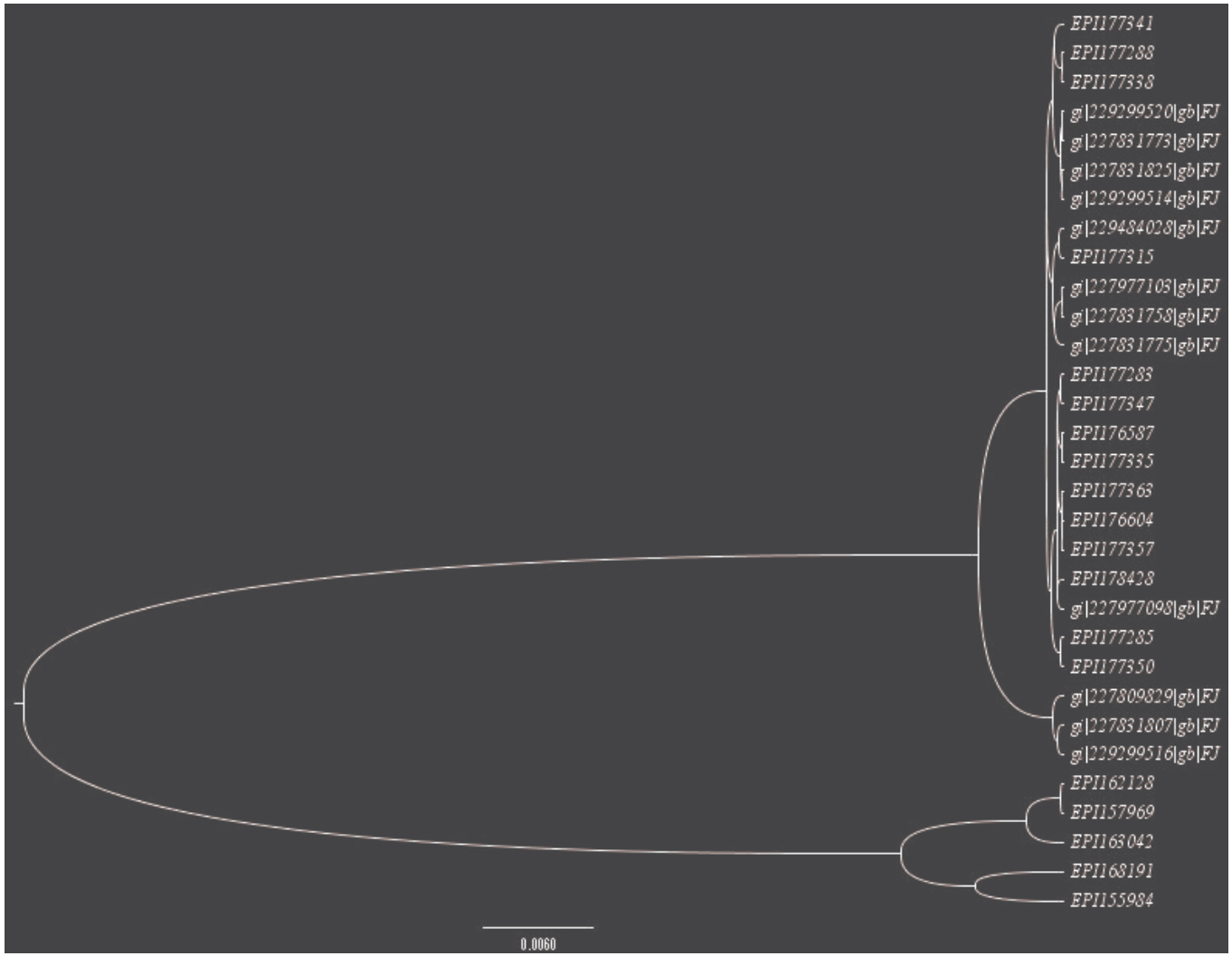
$4Nu = 5$



Conclusiones: coalescencia

- La teoría de coalescencia permite hacer inferencias usando hipótesis nulas *ad hoc*
- Estas inferencias incluyen estimaciones de parámetros y de procesos poblacionales históricos
- El campo se está moviendo hacia generar hipótesis usando simulaciones con el conocimiento de la historia natural de las especies estudiadas

Sobre la evolución y el origen del AH1N1



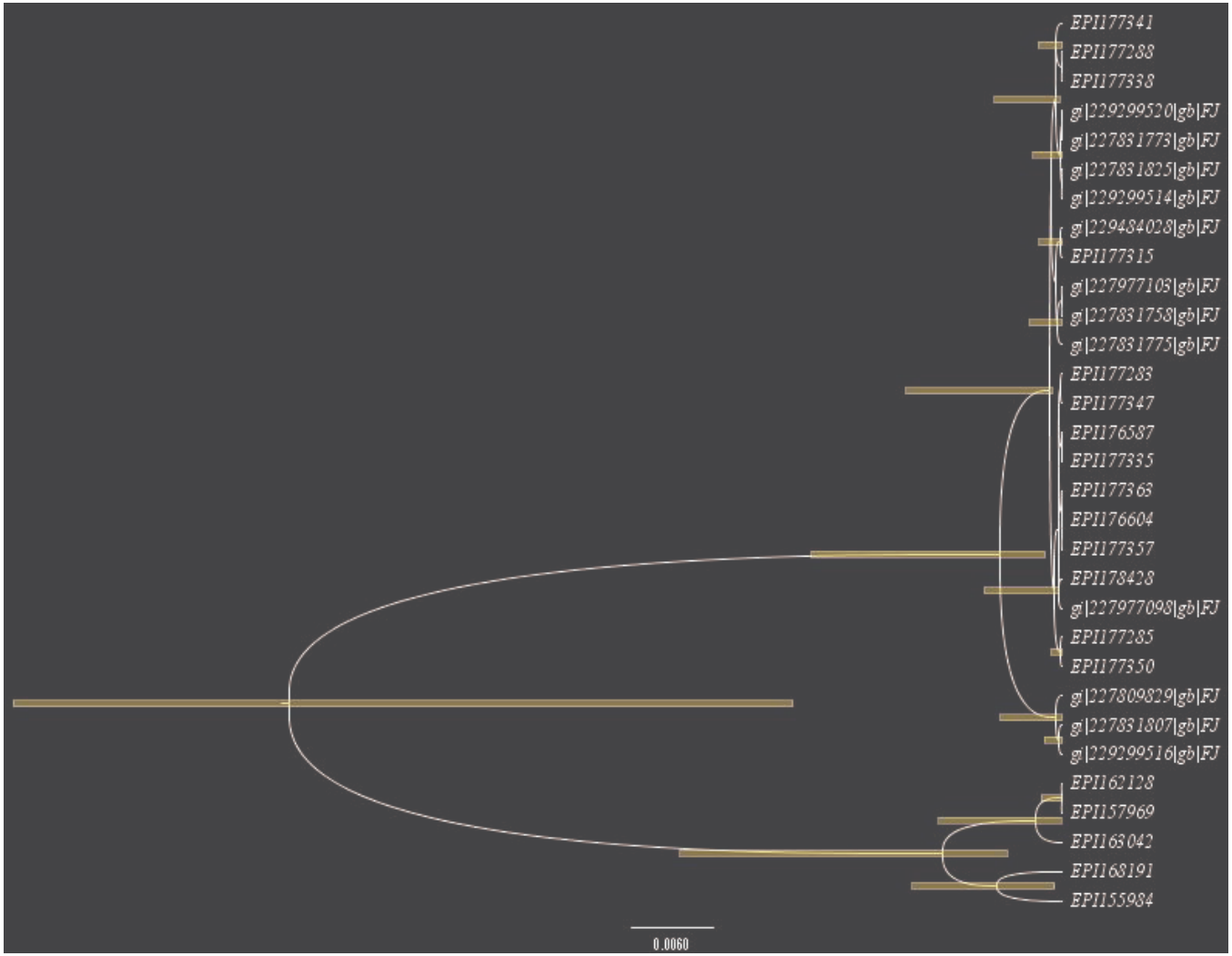
Mexico,
Texas,
NY

California

México 2007 y 2008

BEAST: Bayesian Evolutionary Analysis by Sampling Trees

La topología no está tan bien resuelta



Mexico,
Texas,
NY

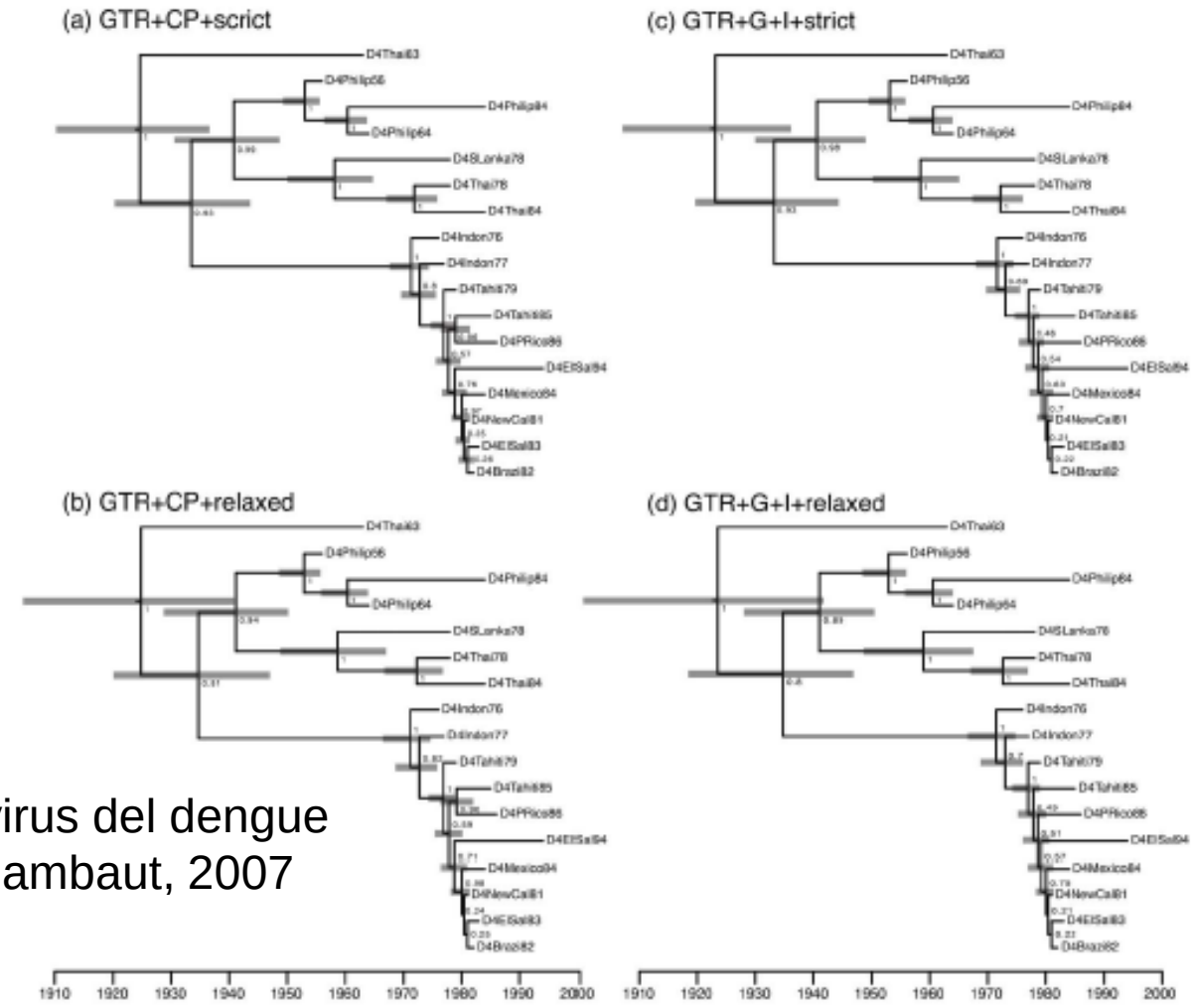
California

México 2007 y 2008

Las barras muestran el 95% de la densidad posterior para cada tiempo de divergencia

5 meses

Se pueden usar distintos modelos de evolución y tasas constantes o no



Evolución del virus del dengue
Drummond y Rambaut, 2007