Un test de hipótesis sobre ancestría

Gabriel Illanes

Centro de Matemática
Facultad de Ciencias
Universidad de la República

15 de octubre de 2019

Información genética de individuos

- Cada individuo tiene 23 pares de cromosomas, de los cuales 22 son no sexuales. Todos los cromosomas son de distinto largo.
- Uno de los cromosomas de ellas es heredada de la madre, y la otra del padre.
- Durante la meiosis, los cromosomas se *recombinan* para crear un nuevo cromosoma, el cual es heredado por su descendencia.
- Los cromosomas se miden en *Morgans*. Es una distancia construida a partir de las frecuencias de recombinación. En una secuencia de bases de largo 1 Morgan, se espera encontrar una recombinación.

Coloreo de los cromosomas

Es posible relacionar la información genética de un individuo con un conjunto de poblaciones ancestrales. Se puede descomponer cada cromosoma no sexual en trozos (haplotipos) donde cada uno de ellos está relacionado a una población ancestral.
En nuestro caso, las poblaciones ancestrales de referencia son Europea, Africana y Nativa Americana.
Un individuo es *puro* relativo a una población ancestral si toda su información genética corresponde a dicha población ancestral.

El problema

Dado el coloreo de los cromosomas de un individuo, queremos realizar un test de hipótesis:

- H_0: El individuo i tiene, al menos, un ancestro nativo americano puro t generaciones en el pasado,
- H_1: no H_0,

donde *ancestro nativo americano puro* es un ancestro cuya información genética corresponda en su totalidad a ancestría nativa americana (o sea, un solo color).

Setting del problema: Datos de Urugenomes

- Contamos, para 20 individuos, con el coloreo de sus 22 cromosomas no sexuales.
- Dentro de los 20 individuos, 10 de ellos declaran tener algún ancestro nativo americano, y 10 declaran tener algún ancestro africano (no necesariamente puros).

Setting del problema: Modelo matemático

- Aproximamos cada cromosoma por un intervalo (continuo), donde su largo está dado por la medida del cromosoma en Morgans.
- No hacemos distinción por sexos.
- "Modelo biológico" de meiosis.
- El proceso en cada cromosoma es independiente del resto de los cromosomas.
- Los 2^t ancestros se cruzan hasta tener descendencia en la generación actual.

Estrategia

La estrategia consiste en 2 pasos:

1. Conseguir "scores" para cada cromosoma por separado. Tenemos que poder estimar la distribución de cada score.
2. Combinar la información de todos los cromosomas en un único p-valor.

Paso 1

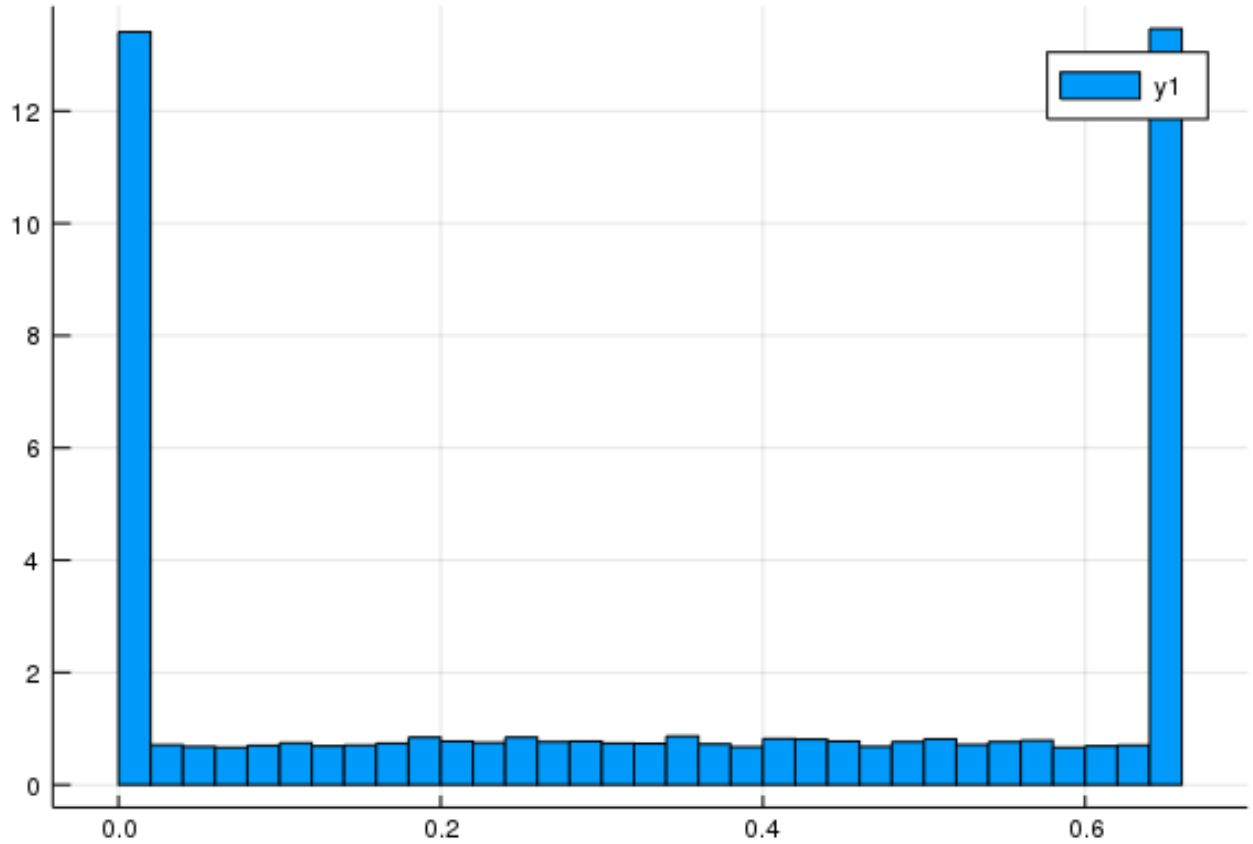
Bajo este setting no markoviano, debemos simular las densidades de los estadísticos que querramos usar.
Para poder simular, reducimos H_0 a un caso borde:

- $\hat{H}_0$: El individuo i tiene, exactamente, un ancestro nativo puro t generaciones en el pasado,
- $\hat{H}_1$: Ningún ancestro es nativo puro t generaciones en el pasado.

Pero necesitamos un estadístico que sea dominado estocásticamente en ese caso borde.

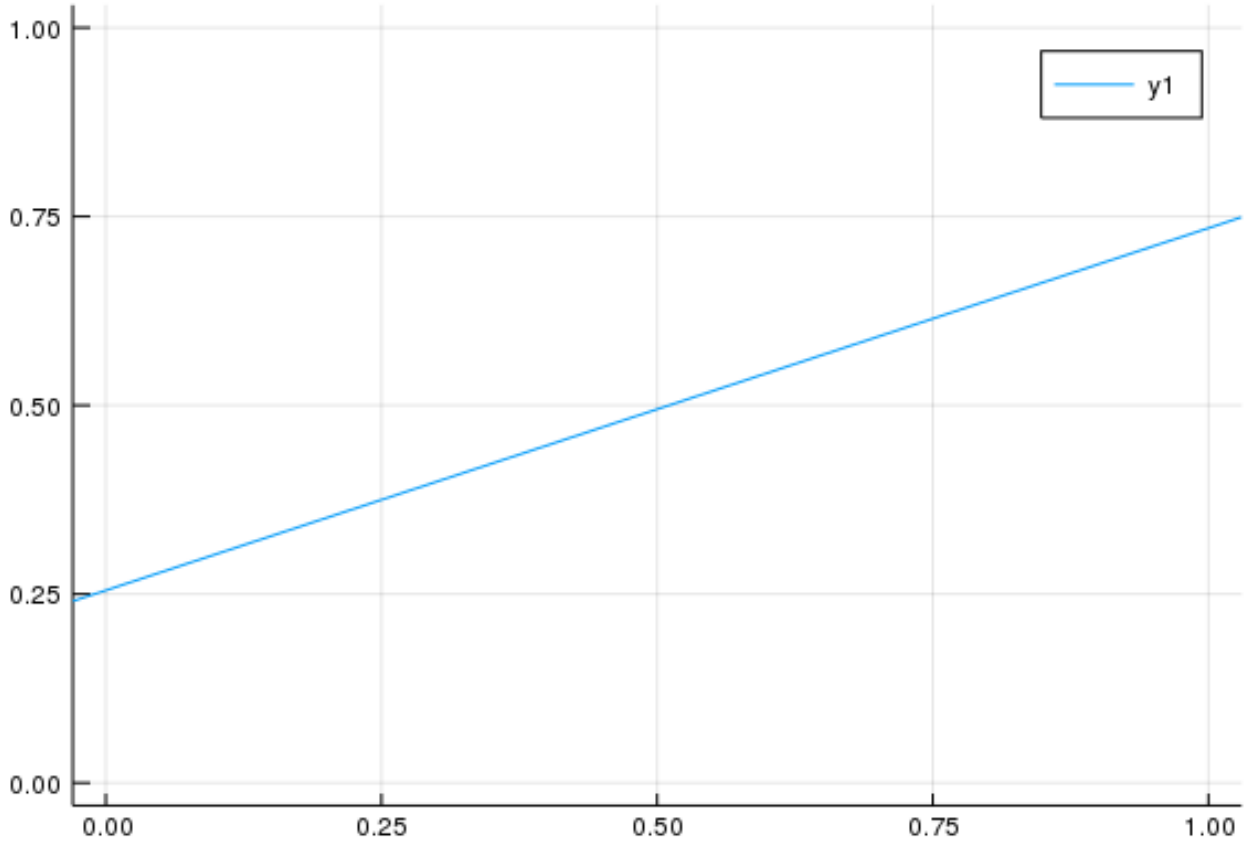
- Largo máximo de los haplotipos nativos,
- Suma de los largos de los haplotipos nativos.

Por ejemplo, para largo máximo l_c...



Podemos estimar la distribución de l_c, F_c mediante simulaciones.

Para obtener estadísticos comparables en todos los cromosomas...



$$p_c = \mathbb{P}(p \leq F_c(l_c) \,|\, l_c > 0)$$

Paso 2

Ya tenemos p_c para $c = 1, \ldots, 22$, y conocemos sus distribuciones.
Hay que combinarlos en un solo estadístico y calcular la región crítica.
Dos opciones:

- Máximo de los p_c (conocemos la distribución de dicho producto).
- Suma de los p_c (también hay que simular la distribución de la suma).

Resultados

Escenario 1: doble de información nativo americana que en H_0 distribuido entre todos los ancestros.

Method	$t = 2$	$t = 3$	$t = 4$	$t = 5$
p_{mm}	0.995	0.952	0.431	0.008
p_{ms}	1.0	0.949	0.004	0.0
p_{sm}	0.996	0.919	0.2	0.001
p_{ss}	1.0	0.794	0.001	0.0

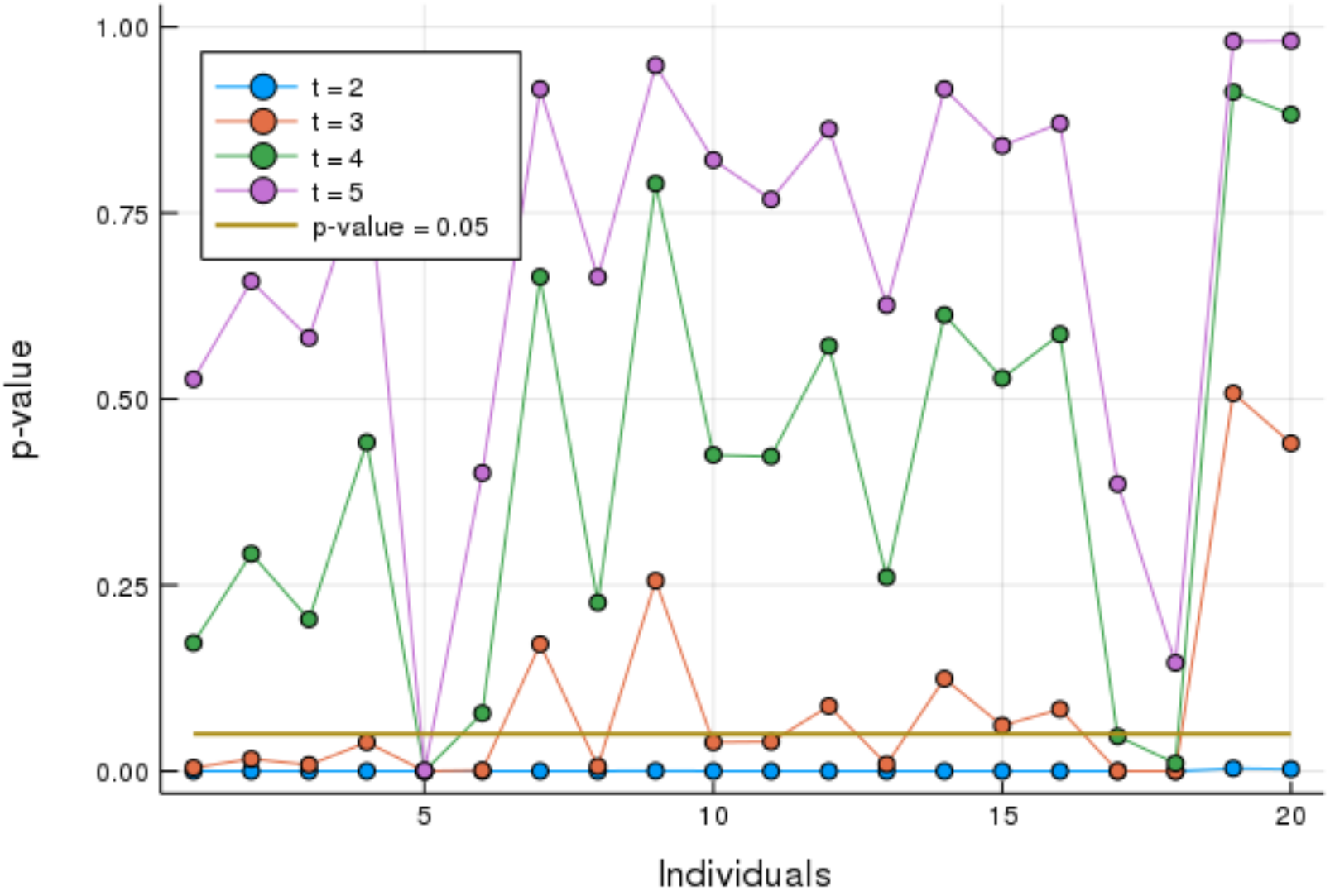
Escenario 2: la mitad (en promedio) de los cromosomas del ancestro nativo americano puro son cambiados por cromosomas puros europeos.

Method	$t = 2$	$t = 3$	$t = 4$	$t = 5$
p_{mm}	0.694	0.504	0.332	0.292
p_{ms}	0.999	0.832	0.574	0.377
p_{sm}	0.701	0.527	0.34	0.249
p_{ss}	0.999	0.823	0.575	0.359

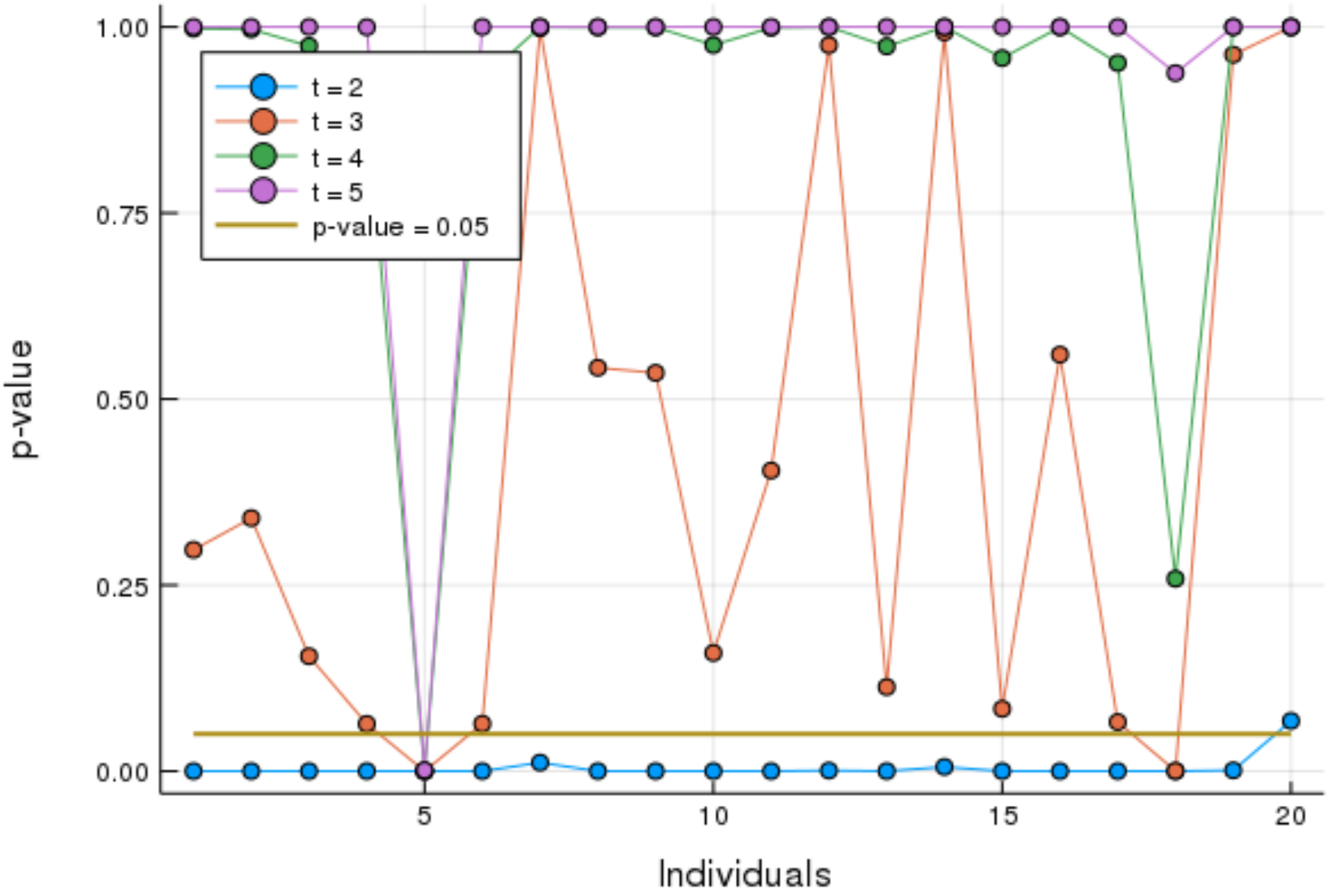
Escenario 3: H_0 cierto para $t+1$.

Method	$t=2$	$t=3$	$t=4$	$t=5$
p_{mm}	0.323	0.261	0.23	0.208
p_{ms}	0.998	0.881	0.622	0.404
p_{sm}	0.347	0.224	0.213	0.201
p_{ss}	1.0	0.891	0.602	0.406

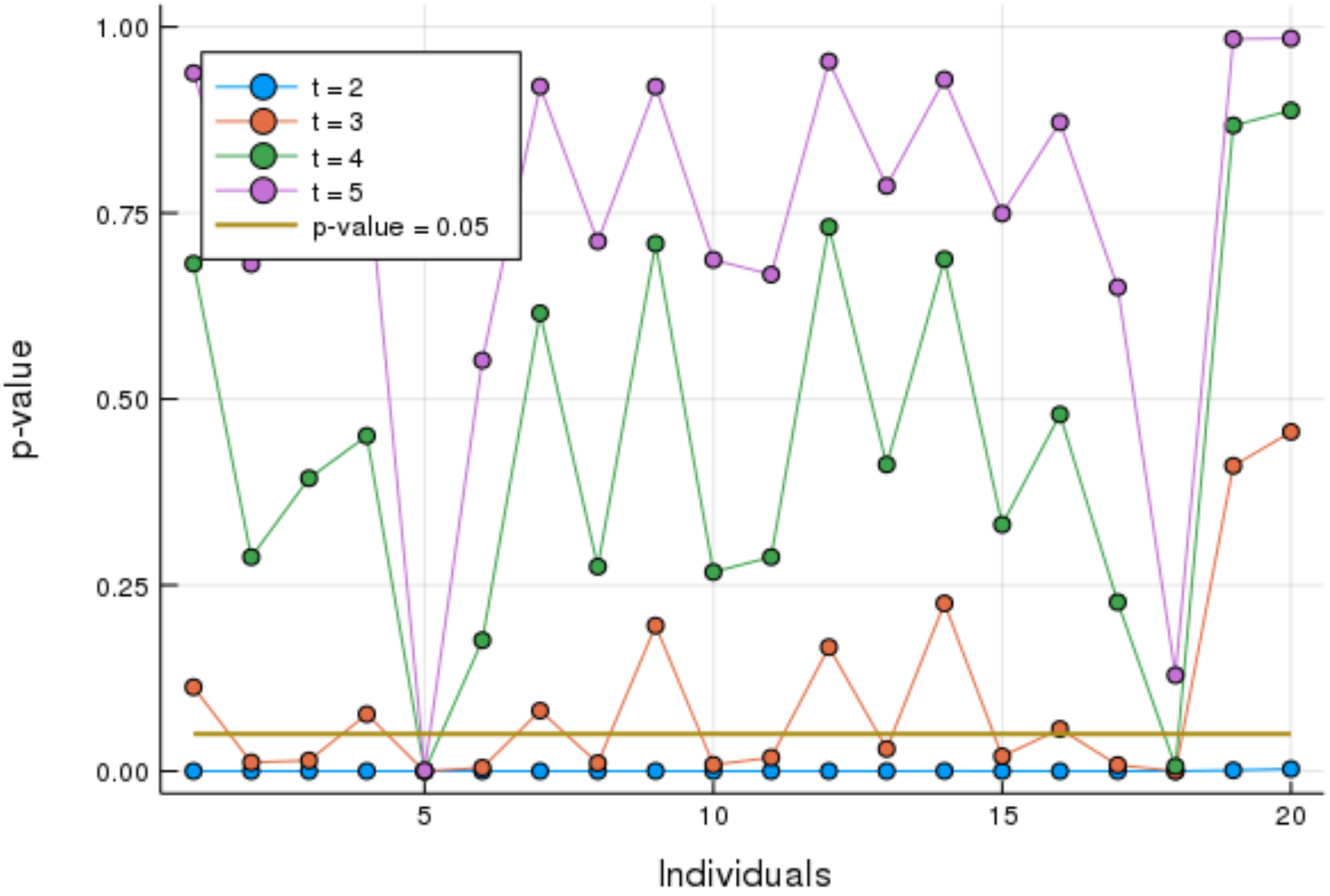
Máximo de largos de haplotipos, máximo de los p_c



Máximo de largos de haplotipos, suma de los p_c



Sumas de largos de haplotipos, máximo de los p_c



Sumas de largos de haplotipos, suma de los p_c