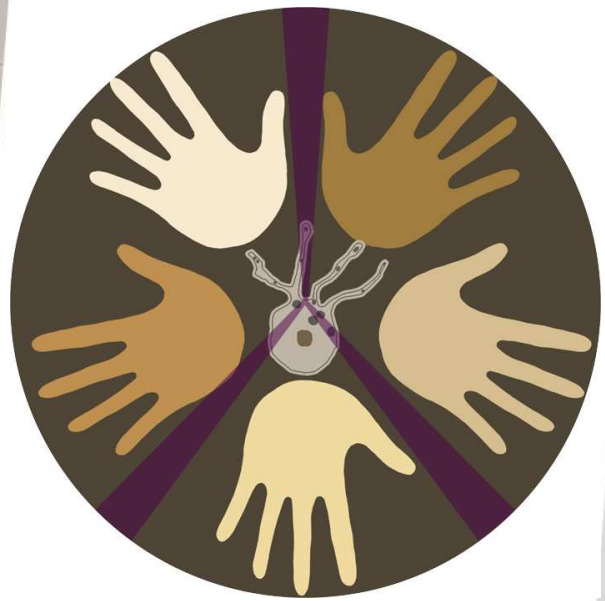
Iniciación a la experimentación en el ámbito de la Biología Celular, Molecular, Genética y Evolutiva

Tema 5. Técnicas de Análisis Poblacional



OpenCourseWare

ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

NAZIOARTEKO BIKAINTASUN CAMPUSA
CAMPUS DE EXCELENCIA INTERNACIONAL

El concepto de diversidad genética

- La diversidad genética es el **número** total de **características genéticas** que existen en cada especie. La diversidad genética es importante porque es la materia sobre la que actúa la selección natural: si no hay diversidad, no hay evolución.
- La **cantidad de diversidad** se determina por factores como:
  - la tasa de mutación (generadora de diversidad)
  - el tamaño de la población (y asociado a éste, la deriva genética)
  - la migración y la selección
  - y, en cierta medida, la recombinación, aunque ésta no crea nueva diversidad sino que reorganiza la ya existente

• Sean:

p_1 la frecuencia de A en la muestra poblacional analizada
p_2 la frecuencia de B
(tal que $p_1 + p_2 = 1$)

• En el equilibrio de Hardy-Weinberg, la frecuencia esperada de un homocigoto es su frecuencia alélica al cuadrado, por lo tanto:

la Homocigosidad esperada es $p_1^2 + p_2^2$

y por consiguiente, la heterocigosidad (diversidad génica) esperada será

1 –Homocigosidad, luego:

Heterocigosidad esperada = $1 - \sum_{i=1}^{n} p_i^2$

O mejor: $n * (1 - \sum_{i=1}^{n} p_i^2) / (n - 1)$

donde n es el número de alelos, en este caso 2.

Las secuencias de ADN y el concepto de haplotipo

- Las **secuencias de ADN** son mucho más **informativas** que los marcadores genéticos clásicos, e incluso que los SNPs independientes.
- Tras secuenciar una región genómica determinada, disponemos de la información de los genomas materno y paterno mezclada y es necesario utilizar técnicas experimentales y/o bioinformáticas para separar ambas informaciones y así saber qué **variantes alélicas** van en el **mismo cromosoma** (i.e. obtener los haplotipos o fases haplotípicas).
- A diferencia de lo que ocurre en el genoma mitocondrial o en la región pseudoautosómica del cromosoma Y, un **locus autosómico** contendrá **dos haplotipos**, que pueden ser iguales o distintos.
- La mayor informatividad de los haplotipos es debida a que la información de la secuencia haplotípica (la ristra de variantes alélicas ligadas), nos permite identificar qué haplotipos son **más antiguos** y cuáles **más modernos**, así como qué haplotipos se derivan de qué otros haplotipos.
- Es decir, el análisis haplotípico nos permite **inferir la genealogía** de los distintos haplotipos observados en un mismo locus o región genómica.

Diversidad en secuencias de ADN

- En los casos de secuencias de ADN, en lugar de hablar de **heterocigosidad**, resulta más informativo hablar de **diversidad génica**
- Bajo la **Teoría Neutral** se asume que, a nivel molecular, la **mayoría de los cambios** o mutaciones son **neutras**, sin ventaja ni desventaja selectiva. Así, se llega a un equilibrio en el que la cantidad de nueva variabilidad que se introduce por mutación, es equilibrada por la que se pierde por deriva debido al tamaño poblacional finito.
- Así, en el equilibrio, la diversidad teórica esperada (θ) es:

$$\theta = 4N_e\mu$$

siendo N_e el tamaño efectivo poblacional y μ la tasa de mutación por locus por generación.

Diversidad genética estimada en locus bialélicos

- Una forma sencilla de analizar la diversidad genética es restringirse al análisis de la **variabilidad de un locus**, o de unos pocos loci.
- Especialmente interesante es el análisis de la diversidad en **loci bialélicos**, que son aquéllos que tienen únicamente 2 alelos (alelos A y B, por ejemplo).
- En los loci bialélicos, una manera de medir su diversidad en una población se basa en estimar la proporción de individuos heterocigotos AB observados en una muestra de esa población.
- Se puede calcular la heterocigosidad esperada (también llamada **diversidad génica**) a partir de las frecuencias de los alelos. A continuación se presentan algunas opciones.

Estimación de la diversidad (θ) a partir de una muestra poblacional

Existen diferentes maneras de estimar θ a partir de un muestreo de individuos de una población, por ejemplo.

1.- A partir del número de sitios segregantes (o polimórficos) **S.**

2.- A partir del número promedio de diferencias nucleotídicas

1.- Estimación de θ a partir del número de sitios segregantes

- Por ejemplo, sea la siguiente muestra de haplotipos:

Haplotipo 1: ACTGGC**T**AAGC**G**CAT**A**CTAG
Haplotipo 2: ACTGGCGAAGCCCATGCTAG
Haplotipo 3: AC**C**GG**T**GAAG**T**CCATGCT**T**G
Haplotipo 4: AC**C**GGCGAAGCCCATGCTAG

- En este caso, S = 7 (el número de lugares en los que se produce alguna variación entre haplotipos).
- Watterson (1975) demostró que cuando una población es panmíctica, está en equilibrio y las mutaciones son neutras, el **valor esperado de S** (E(S)) es: $E(S) = a_1\theta$ en donde $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$ y θ es $4N_e\mu$
- Por lo tanto, podemos despejar y obtener que $\theta = S/a_1$. Así, en nuestro ejemplo anterior, S = 7 y n=4, por lo que a_1 es 1.8333. Consecuentemente una estima de θ (que denominaremos θ_S), será 7/ 1.8333 = 3.818.
- Luego, en este caso, $\theta_S = 3.818$.

2.- Estimación de θ a partir del número promedio de diferencias nucleotídicas entre dos secuencias (k).

Las diferencias nucleotídicas entre dos secuencias (k), se define como:

$$k = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}}{n(n-1)}$$

k_{ij} implica comparar de dos en dos todas las secuencias (i *vs* j) y registrar el número de posiciones nucleotídicas en que difieren, para hallar posteriormente la media de todas las comparaciones realizadas.

En nuestro caso $k_{12} = 3$, $k_{13} = 7$, $k_{14} = 4$, $k_{23} = 4$, $k_{24} = 1$ y $k_{34} = 3$. Por lo tanto la media, k, es 2*(3+7+4+4+1+3)/(4*3) = 3.667.

k es una estima de θ que denominaremos θ_k

k es igual a la suma de las heterocigosidades esperadas en cada sitio nucleotídico

$$k = \sum_{i=1}^{S} h_i$$

Derivación del test D de Tajima (1989)

De acuerdo con lo descrito por Tajima, cuando las mutaciones son selectivamente neutras, la población es panmíctica, está en equilibrio y no hay recombinación, el valor esperado de k es θ. En estas circunstancias, ambas estimas de θ deberían ser aproximadamente iguales.

Sin embargo, cuando hay selección (o actúa otra fuerza evolutiva), lo anterior no se cumple. Así, Tajima diseñó un test que comparaba ambas estimas de θ, y que denominó el test D (de Tajima), que permite detectar la presencia de alguna fuerza selectiva.

$$D = \frac{k - \frac{S}{a_1}}{\sqrt{e_1 S + e_2 S(S-1)}}$$ en donde

$$e_1 = \frac{b_1 - 1/a_1}{a_1} \qquad e_2 = \frac{b_2 - (n+2)/(a_1 n) + a_2/a_1^2}{a_1^2 + a_2} \qquad a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \qquad a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2} \qquad b_1 = \frac{n+1}{3(n-1)} \qquad b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$$

Interpretación del test D de Tajima

La D de Tajima es un estadístico que permite evaluar si los datos (conjunto de secuencias haplotípicas) son consistentes con que la población muestreada evoluciona bajo un modelo neutral (equilibrio mutación-deriva) o no.

- Si la población está en equilibrio mutación-deriva para ese locus (evolución neutral, no selección), el valor de D será estadísticamente indistinguible de cero.

- Si la población está bajo selección sobredominante para ese locus (ventaja del heterocigoto), D será significativamente mayor de cero

- Si la población está bajo selección purificadora para ese locus (eliminación de nuevas mutaciones) o positiva (favorecimiento de una nueva mutación que proporciona mayor fitness), D será significativamente menor de cero

En general, el valor de D depende de cuántos lugares polimórficos detectemos y cómo sean de frecuentes.

Interpretación de D

- Interpretar D puede ser, sin embargo, algo más complicado porque, entre otras cosas, los cambios demográficos afectan también a D. Por ejemplo, si una población ha empezado recientemente a expandirse, podemos obtener una D negativa. Alternativamente, si una población ha sufrido recientemente un cuello de botella, podemos obtener una D positiva.

- Además, el fallo en detectar que D es diferente de cero, puede deberse asimismo a que el tamaño muestral no es lo suficientemente grande como para detectar la señal de la selección.

Diversidad génica e historia evolutiva

- Otro aspecto importante que define los patrones de diversidad genética de las poblaciones es su **historia demográfica**.
- En humanos, nuestra historia evolutiva indica que *Homo sapiens* surgió en África hace unos 200,000 años y que hace unos 100,000 años un pequeño grupo abandonó África y se asentó en el Oriente Próximo, y hace unos 40-50,000 años colonizó Europa y Asia, reemplazando al Neandertal que habitaba en Europa.
- Por lo tanto, es de esperar que, dado el **mayor tamaño poblacional africano** y el **mayor tiempo evolucionando *in situ***, la **diversidad de esta población sea mayor** que la diversidad del mismo locus en Euroasiáticos.
- Cualquier alejamiento de este principio es sospechoso de selección.

Herramientas bioinformáticas para analizar la Diversidad génica

Necesitamos bajarnos 3 programas (free software)

Bioedit	http://www.mbio.ncsu.edu/bioedit/bioedit.html
DnaSP	http://www.ub.edu/dnasp/
Network	http://www.fluxus-engineering.com/sharenet.htm

Otros recursos de internet

SPSmart	http://spsmart.cesga.es/
UCSC genome browser	https://genome-euro.ucsc.edu/

Software propio

SPSmart2haps.exe (execute: c:\dir> SPSmart2haps.exe File-NAME-Root)
(requires a ref file: file-name-root_refseq_hg19.txt)

Previamente, necesitamos saber lo que es un fichero FASTA

Un fichero FASTA es un **fichero de texto simple** que contiene datos de secuencias de ADN (aunque también podría ser de ARN o proteínas) pertenecientes a una conjunto de individuos (nuestra muestra poblacional). Estos datos están representados en un formato denominado a su vez **FASTA**.

Qué es un fichero de texto simple

Es un fichero como el que se genera al usar el "Bloc De Notas" (windows). Contiene texto sencillo, sin negrita, colores etc. Se puede generar también usando Word pero hay que especificar "salvar como texto MS-Dos".

Qué es formato FASTA

En un formato fasta, cada secuencia se representa por dos campos:

a) un identificador de la secuencia
b) la propia secuencia en si

El identificador de secuencia

Es el nombre arbitrario que le damos a la secuencia, precedido por el símbolo "›" y finalizado por un salto de línea (¶).

La secuencia en si

La secuencia haplotípica de ADN, en formato estandard IUB/IUPAC:

A adenina **C** citosina **G** guanina **T** timina

N nucleótido desconocido - normalmente indica una deleción de 1 base

Se aceptan minúsculas y mayúsculas

La secuencia puede ocupar una única línea o varias, dependiendo si decidimos segmentarla secuencia en trozos contiguos de un número de caracteres determinados (normalmente 80). Todas las secuencias deben tener el mismo formato. La secuencia finaliza con un salto de línea (¶).

No se admiten líneas en blanco entre las diferentes secuencias

Ejemplo de formato FASTA

›secuencia1¶
ACTGACGATGACGATACAGTAGCGATGACGATGACGATAGGAGAC
AGCAGACGATATATAGACGATAGCAGTAGACGATACTGACGATGAC
GATACAGTAGCGATGAGATGACGATAGGAGACAGCAGACGCATAT
ATAGACGATAGCAGTAGACGATACTGACT¶
›secuencia2¶
ACTGACGATGACGATACAGTAGCGATGACGATGACGATAGGAGAC
AGCAGACGATATATAGACGATAGCAGTAGAGGATACTGACGATGA
CGATACAGTAGCGATGAGATGACGATAGGAGACAGCAGACGCATA
TGTAGACGATAGCAGTAGACGATACTGACT¶
…

O en segmentos:

›sec1¶
ACTAGCATGACTGTGACGATGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGATGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGATGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
›sec2¶
ACTAGCATGACTGTGACGATGACGATGACGATGAT¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGTTGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGATGTCGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGTTGACGATAGCAG¶
…

NOTA

El símbolo ¶ representa un salto de línea (tecla 'return'); No hay que escribirlo. Sólo lo representamos aquí para visualizar el cambio de línea. Sería visible si abrimos el fichero con Word y especificamos en la sección "Párrafo" que se visualice clicando en el icono correspondiente.

Los ficheros FASTA suelen llevar la extensión .fas (como por ejemplo, en "fichero.fas")

Esquema de Bajada y Formateo de datos (procedimiento detallado en tutoriales 1 y 2)

Vamos a bajarnos y a formatear adecuadamente los datos que necesitamos para el análisis de selección mediante el test D de Tajima (en DnaSP).

1.- Para ello necesitamos definir la región genómica de interés, en este caso será el gen *MC1R*. Averiguamos sus coordenadas genómicas (hg19) mediante el UCSC genome Browser, y nos bajamos la secuencia genómica correspondiente al CDS sólo (no exones UTR). Salvamos como fichero texto plano (MC1R_refseq_hg19.txt). Chequear.

2.- Abrir la página web de SPSmart y, usando las coordenadas anteriores, proceder a la bajada de datos. Bajar los datos correspondientes al conjunto de las poblaciones africanas: (AFR) y de cada una de las poblaciones europeas (cada una por separado, GBR, FIN, CEU, IBS y TS). (procedimiento detallado en tutorial 3)

3.- Ejecutar **SPSmart2haps.exe** con cada población bajada (una vez con AFR, otra con CEU, otra con GBR etc). De los fichero de salida, usaremos de cada el fichero .fas

NOTA: generar dos ficheros a partir de las poblaciones europeas, uno que contenga las poblaciones del Norte de Europa (que denominaremos NEU) y otro que contenga las poblaciones del Sur de Europa (que denominaremos SEU). Usar para ello la propiedad de Merge files de Bioedit. Generaremos así tres ficheros .fas: MC1R_AFR.fas, MC1R_NEU.fas y MC1R_SEU.fas

4.- Abrir cada uno de estos ficheros con DnsSP y calcular la **D de Tajima** (usando el número de sitios segregantes). Calcular la p-value con simulaciones (usando el colaescente). **(procedimiento detallado en tutorial 4)**

Bibliografia

- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*.1989; 123:585-595.

- Watterson GA. Number of segregating sites in genetic models without recombination. *Theor Popul Biol* (1975); 7: 256-276.

Lecturas recomendadas

- Hartl DL & Clark AG. (2007) Principles of population genetics. 4th Ed. Sinauer.

- Joblin M, Hollox E, Hurles M, Kivisild T & Tyler-Smith C. (2014) Human Evolutionary Genetics. 2nd Edition, Garland Science. Pierce B.A. (2008) Genetics. A conceptual approach. 3rd Edition. Freeman and Co.