CARTILLA DE PRÁCTICOS

Curso de Evolución 2024

Laboratorio de Evolución. Facultad de Ciencias. Universidad de la República

Práctico 1

Filogenias

Introducción

Los chimpancé, *Pan troglodytes*, de la región centro-oeste del continente Africano son reconocidos como un reservorio de virus de inmunodeficiencia en Simios (SIVcpzPtt), los cuales han cruzado la barrera específica en al menos dos oportunidades, resultando en la pandemia provocada por el Síndrome de Inmuno Deficiencia Adquirida (HIV-1, grupo M) y por otro lado en la infección aislada de unos pocos individuos en Camerún (HIV-1, grupo N). Un tercer linaje de virus HIV-1 (grupo 0), también de la región centro-oeste de África, cae igualmente dentro de la radiación de los virus de tipo «SIVcpzPtt». Más de 30 especies de primates son portadores de virus que provocan inmunodeficiencia en Simios, pero los chimpancés son los principales portadores de los tipos cercanamente emparentados al HIV-1. Con la finalidad de establecer el origen de la cepa HIV-1 (grupo 0) se secuenciaron algunos genes de varias muestras de chimpancés (SIVcpz) y gorilas (*Gorilla gorilla*, SIVgor) de Camerún. El objetivo de esta actividad consiste, mediante un análisis filogenético de secuencias de ADN, investigar el posible origen y relacionamiento de las diferentes cepas de HIV-1 presentes en chimpancés, gorilas y humanos.

Datos y programa de análisis

La base de datos a analizar consiste en secuencias de los genes env –que codifica proteínas de la envoltura– y pol –que codifica la transcriptasa inversa– del virus HIV-1. Las secuencias ya se encuentran alineadas (las homologías posicionales entre las distintas secuencias ya están establecidas), por lo que ya están listas para ser analizadas.

El programa que se usará para generar las hipótesis filogenéticas es el MEGA11. Éste es un programa que se baja gratis de la red en http://www.megasoftware.net y que tiene varias prestaciones, incluyendo el estudio descriptivo de las secuencias y reconstrucciones filogenéticas mediante métodos de basados en distancias genéticas, máxima parsimonia y máxima verosimilitud.

Actividades a realizar durante el práctico

1) Abrir el archivo “HIV.meg” en MEGA11 (“File\Open A File/Session”). Visualizar la matriz de datos (“Data\Explore Active Data”) e identificar los primeros 5 sitios variables (botón “V”) y los primeros 3 informativos (botón “Pi”). ¿Por qué algunos de los sitios variables no son informativos?

2) Ir a la ventana principal de MEGA11 y realizar un análisis utilizando el criterio de Máxima Parsimonia [“Phylogeny\Construct/Test Maximum Parsimony Tree(s)”]. Analice el apoyo de los clados obtenidos utilizando «bootstrap», con 100 pseudoréplicas (opción “No. of Bootstrap Replication -> 100”). Mantenga los demás parámetros del análisis en sus condiciones por defecto. Luego de obtenido el árbol filogenético, moverse de la pestaña “Original Tree” a la opción “Bootstrap consensus tree”. ¿Qué representan las

reconstrucciones filogenéticas que se muestran en cada una de esas dos opciones? Definir el grupo externo en el “Bootstrap consensus tree”, utilizando las secuencias virales obtenidas del Mono verde Africano, *Cercopithecus aethiops* (SIVagm). Para lograr esto, seleccione primero la rama que representa a ese taxón y luego la opción: “Subtree\Root” (también se puede hacer desde un botón ubicado en la barra lateral). ¿Qué función cumple el grupo externo?

3) Registre la longitud e índice de consistencia del árbol consenso obtenido. Esta información se encuentra disponible en la leyenda de la figura así como en la opción “i” del menú. Registre el valor de apoyo de los clados obtenidos utilizando bootstrap. Indique cuáles son los 3 clados que reciben menor apoyo estadístico.

4) Compare los valores de bootstrap con los obtenidos por otros compañeros. ¿Por qué difieren?

5) Opcional: Repita la reconstrucción filogenética –incluyendo el análisis de bootstrap– utilizando el algoritmo de unión de vecinos (“Phylogeny\Construct/Test Neighbor-Joining Tree...”).

6) De acuerdo a los resultados obtenidos con la reconstrucción filogenética discutir el posible origen y vías de contagio interespecíficas de los diferentes grupos de HIV-1. ¿Cuál fue el origen del contagio del grupo “HIV-1 (grupo 0)” en humanos? Teniendo en cuenta las vías de transmisión del virus, ¿cuáles serían las posibles formas de que se traspasaron las barreras específicas?

Basado en:
Hillis, D. 2010. Phylogenetic Progress and Applications of the Tree of Life, 421-449 p. En: Evolution since Darwin: The first 150 years, Editado por: Bell, M.; Futuyma, D.; Eanes, W.; & Levinton, J.; 688pp.
Van Heuverswyn et al. 2006. Human immunodeficiency viruses: SIV infection in wild gorillas. Nature 444:164, doi:10.1038/444164a.

Práctico 2

Métodos Comparativos Filogenéticos

Resumen del problema

Es bien sabido que la presión parcial de oxígeno disminuye con la altura. Las personas no habituadas “se apunan” o sienten malestar al trasladarse a zonas de montaña, especialmente al momento de realizar actividades físicas exigentes. Un individuo que se traslada desde una zona baja hasta una de altura, experimenta a lo largo del tiempo una serie de ajustes fisiológicos para paliar al menos parte de esos desajustes.

Por lo tanto, cabe preguntarse si hay cambios genéticos que han sido favorecidos por la selección natural y que permiten a especies y poblaciones que viven a altas elevaciones adaptarse mejor a dichas condiciones. Tanto en humanos como en especies animales (y muchas otras), hay estudios orientados a identificar estos cambios.

Puesto que la afinidad de la sangre por el oxígeno es un factor clave para la vida en altura, y dicha afinidad depende fuertemente de las características de la hemoglobina (recordemos que la estructura cuaternaria de la hemoglobina combina dos cadenas de tipo alfa y dos de tipo beta en un tetrámero, en torno a un núcleo de hierro), esta proteína ha sido el blanco de muchos estudios.

Natarajan et al. (2016) se plantean identificar algunas de las adaptaciones de la hemoglobina para la vida en la altura. La hipótesis de trabajo es que la selección natural pudo haber favorecido cambios en las características de la hemoglobina de las especies asociados a la elevación en la que vive cada una.

Algunas de las ideas del artículo son: - Realizar un estudio comparando múltiples especies de aves, procurando elegir pares de especies cercanamente relacionadas, de modo que una de las especies de cada par viva en tierras altas y otra en las tierras bajas cercanas. - Para cada una de estas especies, aislar la hemoglobina y estudiar su afinidad con el oxígeno en el laboratorio.

El estudio incluye un análisis de los cambios en las secuencias de las hemoglobinas, que usaremos más adelante en el curso. Por el momento, extraemos del artículo los siguientes datos para cada una de las 56 especies estudiadas:

1. La elevación en la que viven (específicamente la altura de la localidad en la que fueron estudiadas).
2. La afinidad de su alfa globina por el oxígeno, tomada en condiciones controladas de laboratorio (en un medio que aproxima las condiciones en sangre), resumida por el valor conocido como P50.
3. El árbol filogenético que se utiliza en el artículo. Este árbol se toma como la mejor hipótesis disponible de las relaciones entre las especies.

Preguntas 1

¿Cómo debería cambiar el P50 para que la hemoglobina tenga mayor afinidad con el oxígeno, por ejemplo en una especie de altura?

Activando los paquetes de R

Para esta actividad práctica usaremos, R (https://cran.r-project.org/) el cual es un entorno y lenguaje de programación enfocado al análisis estadístico. Se recomienda el uso RStudio (https://rstudio.com/), que es un entorno de desarrollo integrado (IDE), en conjunción con R, lo cual facilita el manejo de datos y la realización de los análisis.

Utilizaremos los siguientes paquetes de R, los cuales deben ser previamente instalados:

tidyverse: una colección de paquetes que facilitan el análisis de datos. *broom*: para formatear salida de modelos de ajuste de estos datos *phytools*: varias funciones para análisis filogenéticos, principalmente orientado a la biología comparada.

Trabajando con los datos originales

Mediante el comando “setwd(dir)” podemos indicarle a R en donde están ubicados los archivos con los cuales vamos a estar trabajando, siendo “dir” la ruta hacia nuestro directorio (e.g. C:\User\Desktop\R).También podemos ubicar el directorio de trabajo de R mediante “getwd()” y colocar ahí nuestros archivos de interés.

El bloque de código siguiente lee solamente los datos de elevación y los de P50 de la HbA.

```
# activamos paquetes
library("phytools")
library("tidyverse")
library("formattable")

# Leemos los datos, tenemos pares de spp. en distinta elevación (categoria)
# más la elevación ocupada + P50 de cada spp

Datos <- read_tsv("Datos.tsv")
## Rows: 56 Columns: 5
## ── Column specification ──────────────────────────────────────────────────────
## Delimiter: "\t"
## chr (3): Familia, Especie, Elev.cat
## dbl (2): Elevacion, P50
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
# notar cuantas columnas tengo y que tipo de variable es (categorica y numérica)
Datos # veo rapidamente los datos
## # A tibble: 56 × 5
## Familia Especie Elevacion P50 Elev.cat
## <chr> <chr> <dbl> <dbl> <chr>
## 1 Columbidae Metriopelia_melanoptera 4178 26.9 Alta
## 2 Columbidae Columbina_cruziana 372 28.4 Baja
```

```
## 3 Caprimulgidae Hydropsalis_longirostris 4401 30.2 Alta
## 4 Caprimulgidae Hydropsalis_decussata 309 36.1 Baja
## 5 Trochilidae Colibri_coruscans 4030 31.1 Alta
## 6 Trochilidae Schistes_geoffroyi 1395 36.8 Baja
## 7 Trochilidae Selasphorus_platycercus 2470 38.2 Alta
## 8 Trochilidae Archilochus_alexandri 2050 39.1 Baja
## 9 Trochilidae Amazilia_viridicauda 3005 24.2 Alta
## 10 Trochilidae Amazilia_amazilia 366 29.8 Baja
## # ℹ 46 more rows
Datos %>% count(Elev.cat) # un conteo de ocurrencias en la variable categoría de elevación
## # A tibble: 2 × 2
## Elev.cat n
## <chr> <int>
## 1 Alta 28
## 2 Baja 28
```

Pregunta 2

La tabla de datos incluye varios pares de especies de un mismo género. Elegir algunos de esos pares para discutir:

1. ¿De qué especies se trata? Averiguar algo de los nombres comunes, familias a las que pertenecen, distribución geográfica.
2. ¿Qué tendencias se observan al examinar en varios de esos pares la relación entre altura y P50?

Examinando la relación entre elevación y P50

En el siguiente bloque se obtiene el coeficiente de correlación entre las dos variables de interés, se aplica un modelo de regresión lineal entre dichas variables, y se grafican los valores junto con la línea de tendencia obtenida en la regresión.

```
# Análisis exploratorios
class(Datos) # que tipo de objeto es? Puede ser vector, list, matrix, data frame, etc...
## [1] "spec_tbl_df" "tbl_df" "tbl" "data.frame"
# Datos %>% formattable() # imprimo en pantalla toda la tabla formateada
summary(Datos) # Resumen de número y tipo de variables y nro de observaciones
## Familia Especie Elevacion P50
## Length:56 Length:56 Min. : 39.0 Min. :17.07
## Class :character Class :character 1st Qu.: 370.5 1st Qu.:27.77
## Mode :character Mode :character Median :2774.0 Median :31.67
## Mean :2548.0 Mean :32.04
## 3rd Qu.:4318.8 3rd Qu.:37.58
## Max. :4800.0 Max. :44.69
## Elev.cat
## Length:56
## Class :character
## Mode :character
# cual es la media de P50 y Elevación en este dataset
```

```
summarise(Datos, mean(P50)) # veo un estadístico de una variable en particular
## # A tibble: 1 × 1
## `mean(P50)`
## <dbl>
## 1 32.0
summarise(Datos, mean(Elevacion)) # veo un estadístico de una variable en
particular
## # A tibble: 1 × 1
## `mean(Elevacion)`
## <dbl>
## 1 2548.
Datos %>% count(Familia) # cuantas observaciones de cada Familia
## # A tibble: 10 × 2
## Familia n
## <chr> <int>
## 1 Anatidae 16
## 2 Caprimulgidae 2
## 3 Columbidae 2
## 4 Emberizidae 2
## 5 Fringillidae 2
## 6 Furnariidae 2
## 7 Hirundinidae 2
## 8 Thraupidae 8
## 9 Trochilidae 18
## 10 Troglodytidae 2


Datos %>% arrange(desc(Elevacion)) # ordenamos de mayor a menor, por Elevación
## # A tibble: 56 × 5
## Familia Especie Elevacion P50 Elev.cat
## <chr> <chr> <dbl> <dbl> <chr>
## 1 Anatidae Lophonetta_s_alticola 4800 25.1 Alta
## 2 Anatidae Chloephaga_melanoptera 4700 27.6 Alta
## 3 Anatidae Anas_georgicaH 4600 35.6 Alta
## 4 Anatidae Anas_c_orinoma 4600 29.4 Alta
## 5 Anatidae Anas_puna 4600 27.3 Alta
## 6 Trochilidae Aglaeactis_castelnaudii 4578 17.2 Alta
## 7 Anatidae Merganetta_armataH 4500 26.6 Alta
## 8 Hirundinidae Notiochelidon_murina 4470 30.9 Alta
## 9 Caprimulgidae Hydropsalis_longirostris 4401 30.2 Alta
## 10 Furnariidae Cinclodes_albiventris 4401 25.1 Alta
## # ℹ 46 more rows
Datos %>% arrange(P50) # ordenamos x P50
## # A tibble: 56 × 5
## Familia Especie Elevacion P50 Elev.cat
## <chr> <chr> <dbl> <dbl> <chr>
## 1 Troglodytidae Troglodytes_aedonH 4375 17.1 Alta
## 2 Trochilidae Aglaeactis_castelnaudii 4578 17.2 Alta
## 3 Trochilidae Heliodoxa_leadbeateri 1890 17.2 Baja
## 4 Trochilidae Coeligena_violifer 3779 19.1 Alta
## 5 Trochilidae Oreotrochilus_estella 4391 20.2 Alta
```

```
## 6 Trochilidae Coeligena_coeligena 2132 22.9 Baja
## 7 Trochilidae Amazilia_viridicauda 3005 24.2 Alta
## 8 Furnariidae Cinclodes_albiventris 4401 25.1 Alta
## 9 Anatidae Lophonetta_s_alticola 4800 25.1 Alta
## 10 Troglodytidae Troglodytes_aedonL 143 25.9 Baja
## # ℹ 46 more row
# ahora podemos encadenar alguno de estos comandos
# ¿cual es la media en las dos categorñías de elevacion?
Datos %>% group_by(Elev.cat) %>% summarise(mean(Elevacion))
## # A tibble: 2 × 2
## Elev.cat `mean(Elevacion)`
## <chr> <dbl>
## 1 Alta 4042.
## 2 Baja 1054.
hist(Datos$Elevacion)
hist(Datos$P50)
# dentro de una familia de interés
Datos %>% filter(Familia == "Trochilidae") %>% summarise(mean(P50))
## # A tibble: 1 × 1
## `mean(P50)`
## <dbl>
## 1 29.3
Datos %>% filter(Familia == "Trochilidae") %>% group_by(Elev.cat) %>%
summarise(mean(Elevacion))
## # A tibble: 2 × 2
## Elev.cat `mean(Elevacion)`
## <chr> <dbl>
## 1 Alta 3826.
## 2 Baja 1936.
Datos %>% filter(Familia == "Trochilidae") %>% group_by(Elev.cat) %>%
summarise(mean(P50))
## # A tibble: 2 × 2
## Elev.cat `mean(P50)`
## <chr> <dbl>
## 1 Alta 27.2
## 2 Baja 31.5


# Exploro relación P50 vs. Elevación (categórica)

ggplot(Datos, aes(Elev.cat, P50)) + # boxplt
geom_boxplot(outlier.colour = "red", outlier.shape = 1) + # vemos los outliers
en rojo
theme_classic() + ggtitle("P50 versus Elevación") + xlab("Elevación")
ggplot(Datos, aes(Elev.cat, P50)) +
geom_boxplot(outlier.colour = "red", outlier.shape = 2) +
geom_jitter(width = 0.5) + # hacemos esto solo para ver mejor todos los valores
theme_classic() + ggtitle("P50 versus Elevación") + xlab("Elevación")
# Ahora vemos la relación P50 vs. Elevación (numérica)
# Calculamos el coeficiente de correlación entre las dos variables
R = cor(Datos$Elevacion,Datos$P50)
```

```r
# Aplicar un modelo lineal (lm) de regresión. Se define a P50 como la variable
dependiente
# y a Elevacion como la variable independiente.

ggplot(Datos, aes(Elevacion, P50)) + geom_point() +
geom_smooth(method='lm') + theme_classic() + ggtitle("Relación entre P50 de HbA
y elevación")
## `geom_smooth()` using formula = 'y ~ x'
Regresion <- lm(Datos$P50~Datos$Elevacion)
summary.lm(Regresion)
##
## Call:
## lm(formula = Datos$P50 ~ Datos$Elevacion)
##
## Residuals:
## Min 1Q Median 3Q Max
## -15.8125 -4.1662 0.8666 4.4904 12.3688
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.9080738 1.5388246 23.335 < 2e-16 ***
## Datos$Elevacion -0.0015162 0.0004991 -3.038 0.00367 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.483 on 54 degrees of freedom
## Multiple R-squared: 0.1459, Adjusted R-squared: 0.1301
## F-statistic: 9.228 on 1 and 54 DF, p-value: 0.003667
# como alternativa podemos obtener una tabla ya formateada o lista
# broom::tidy(Regresion) %>% formattable() # tabla formateada
# broom::glance(summary(Regresion)) %>% %>% formattable() # tabla formateada

# Graficamos los valores de P50 y elevación, agregando la línea de ajuste
# ...obtenida de la regresión lineal.

#

Datos %>% filter(Familia == "Anatidae" | Familia == "Trochilidae" ) %>%
ggplot( aes(Elevacion, P50, group=Familia)) + geom_point(aes(color= Familia)) +
geom_smooth(method='lm') + theme_classic() +
ggtitle("Relación entre P50 de HbA y elevación ")
## `geom_smooth()` using formula = 'y ~ x'
```

Pregunta 3

1. ¿Cuál es el valor del coeficiente de correlación? ¿Qué sugiere respecto a la relación entre las variables (signo, intensidad)?

2. Examinar la tabla de regresión. ¿Qué fracción de la varianza está explicada por el modelo?

3. Identificar y explicar los parámetros relevantes del modelo y el p-valor asociado a cada uno. ¿Qué sugieren estos resultados sobre la hipótesis de trabajo?

4. Examinar la gráfica, incluyendo la predicción obtenida por regresión lineal. ¿Qué sugiere la gráfica y cómo se relaciona con los puntos discutidos más arriba?

Las especies, y por lo tanto sus rasgos, no pueden ser tratadas como variables independientes desde el punto de vista estadístico dado que comparten una historia evolutiva en común.Este problema se hace aún más evidente en taxa cercanamente emparentados. La respuesta a tal problema fue propuesta por Felsenstein (1985) mediante el cálculo de **Contrastes Filogenéticos Independientes (PIC por sus siglas en inglés)**, mediante el cual se transforma a los rasgos analizados en variables independientes, empleando la filogenia de las especies como marco de análisis. En la siguiente sección, vamos a abordar este problema y analizaremos nuevamente la correlación entre la P50 y la Altura, pero esta vez corrigiendo la falta de independencia de los datos, utilizando los PIC como nuevas variables, independientes de la historia evolutiva compartida entre las diferentes especies de aves.

```
#leemos el árbol filogenético
tree<-read.tree("tree.tre")

#enraizamos el árbol en su punto medio
treeR<-midpoint.root(tree)

#visualizamos el árbol enraizado
plotTree(treeR, edge.width=1, ftype="i", fsize=0.7)
# Para simplificar los comandos, leemos otro archivo que contiene solo las
especies y las máximas alturas registradas,
# y leemos otro archivo con los valores de P50
elevacion_2<-read.table("Elevacion.txt")
P50_2<-read.table("HbA_KCLHIP_P50.txt")

#Creamos un vector numérico de las variables para calcular los contrastes
VP50 <- P50_2$P50
Velevacion <- elevacion_2$Elevacion

#Paso necesario para que los datos y los taxa terminales se asocien
correctamente
names(VP50) <- row.names(P50_2)
names(Velevacion) <- row.names(elevacion_2)

enframe(VP50) # acá corroboramos como asociamos nombre spp. ~ valor
## # A tibble: 56 × 2
```

```
## name value
## <chr> <dbl>
## 1 Metriopelia_melanoptera 26.9
## 2 Columbina_cruziana 28.4
## 3 Hydropsalis_longirostris 30.2
## 4 Hydropsalis_decussata 36.1
## 5 Colibri_coruscans 31.1
## 6 Schistes_geoffroyi 36.8
## 7 Selasphorus_platycercus 38.2
## 8 Archilochus_alexandri 39.1
## 9 Amazilia_viridicauda 24.2
## 10 Amazilia_amazilia 29.8
## # ℹ 46 more rows
enframe(Velevacion) # para la otra variable
## # A tibble: 56 × 2
## name value
## <chr> <int>
## 1 Metriopelia_melanoptera 4178
## 2 Columbina_cruziana 372
## 3 Hydropsalis_longirostris 4401
## 4 Hydropsalis_decussata 309
## 5 Colibri_coruscans 4030
## 6 Schistes_geoffroyi 1395
## 7 Selasphorus_platycercus 2470
## 8 Archilochus_alexandri 2050
## 9 Amazilia_viridicauda 3005
## 10 Amazilia_amazilia 366
## # ℹ 46 more rows
#calculamos los contrastes filogenéticamente independientes
ContrasteVP50 <- pic(VP50, treeR)
ContrasteVelevacion <- pic(Velevacion, treeR)

#si queremos extraer además de los contrastes, su varianza asociada
ContrasteVP50.var <- pic(VP50, treeR, var.contrasts=TRUE)
ContrasteVelevacion.var <- pic(Velevacion, treeR, var.contrasts=TRUE)

#Ahora visualizamos los contrastes filogenéticos calculados anteriormente
# ContrasteVP50
enframe(ContrasteVP50) # aca lo veo como 'tabla'
## # A tibble: 55 × 2
## name value
## <chr> <dbl>
## 1 57 1.22
## 2 58 1.56
## 3 59 -1.81
## 4 60 -4.43
## 5 61 6.93
## 6 62 5.38
## 7 63 1.63
## 8 64 -2.49
## 9 65 -1.63
```

```
## 10 66 0.771
## # ℹ 45 more rows
#ContrasteVelevacion
enframe(ContrasteVelevacion) # aca lo veo como 'tabla'
## # A tibble: 55 × 2
## name value
## <chr> <dbl>
## 1 57 -68.2
## 2 58 443.
## 3 59 301.
## 4 60 -712.
## 5 61 -58.0
## 6 62 196.
## 7 63 353.
## 8 64 -434.
## 9 65 79.9
## 10 66 -626.
## # ℹ 45 more rows
# vemos los contrastes filogenéticos calculados y también su varianza asociada

# ContrasteVP50.var
# ContrasteVelevacion.var

enframe(ContrasteVP50.var) # la varianza P50
## # A tibble: 55 × 2
## name value[,"contrasts"] [,"variance"]
## <chr> <dbl> <dbl>
## 1 57 1.22 2.93
## 2 58 1.56 4.66
## 3 59 -1.81 3.32
## 4 60 -4.43 3.32
## 5 61 6.93 3.32
## 6 62 5.38 3.32
## 7 63 1.63 3.32
## 8 64 -2.49 3.31
## 9 65 -1.63 3.25
## 10 66 0.771 3
## # ℹ 45 more rows
enframe(ContrasteVelevacion.var) # la varianza Elevacion
## # A tibble: 55 × 2
## name value[,"contrasts"] [,"variance"]
## <chr> <dbl> <dbl>
## 1 57 -68.2 2.93
## 2 58 443. 4.66
## 3 59 301. 3.32
## 4 60 -712. 3.32
## 5 61 -58.0 3.32
## 6 62 196. 3.32
## 7 63 353. 3.32
## 8 64 -434. 3.31
## 9 65 79.9 3.25
```

```
## 10 66 -626. 3
## # ℹ 45 more rows
# Los contrastes asociados a P50 aparecerán en azul y los asociados a la
Elevación
# aparecerán en verde.

plot(treeR, no.margin=TRUE,edge.width=1 ,cex=c(0.6,0.6))
nodelabels(round(ContrasteVelevacion.var [,1], 3), adj = c(1, -0.4), frame="n",
col = "blue", cex=0.6)
#volvemos a plotear la filogenia solo para que nos se superpongan ambos
contrastes
plot(treeR, no.margin=TRUE, edge.width=1 ,cex=c(0.6,0.6))
nodelabels(round (ContrasteVP50.var [,1], 3), adj = c(1, 1.1), frame="n",
cex=0.6, col = "darkgreen")
#Analizamos la evolución correlacionada de ambos rasgos a partir de los archivos
sin el cálculo de varianzas
RegresionP50_elevacion <- lm(ContrasteVP50~ContrasteVelevacion)

#Visualizamos los parámetros de la regresión

summary.lm(RegresionP50_elevacion)
##
## Call:
## lm(formula = ContrasteVP50 ~ ContrasteVelevacion)
##
## Residuals:
## Min 1Q Median 3Q Max
## -9.4792 -1.6509 -0.0061 2.1798 10.1324
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.17539 0.60601 -0.289 0.773392
## ContrasteVelevacion -0.00146 0.00037 -3.947 0.000235 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.671 on 53 degrees of freedom
## Multiple R-squared: 0.2271, Adjusted R-squared: 0.2126
## F-statistic: 15.58 on 1 and 53 DF, p-value: 0.0002347
broom::tidy(summary(RegresionP50_elevacion)) # %>% formattable() # tabla
formateada
## # A tibble: 2 × 5
## term estimate std.error statistic p.value
## <chr> <dbl> <dbl> <dbl> <dbl>
## 1 (Intercept) -0.175 0.606 -0.289 0.773
## 2 ContrasteVelevacion -0.00146 0.000370 -3.95 0.000235
broom::glance(summary(RegresionP50_elevacion)) # %>% formattable() # tabla
formateada
## # A tibble: 1 × 8
## r.squared adj.r.squared sigma statistic p.value df df.residual nobs
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl>
```

```
## 1 0.227 0.213 3.67 15.6 0.000235 1 53 55
broom::glance(summary(Regresion)) # %>% formattable() # tabla formateada
## # A tibble: 1 × 8
## r.squared adj.r.squared sigma statistic p.value df df.residual nobs
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl>
## 1 0.146 0.130 6.48 9.23 0.00367 1 54 56
# broom::glance(summary(Regresion)) %>% formattable() # tabla anterior

#visualizamos la relación entre los contrastes y agregamos una línea de
tendenica a la regresión

#
Contraste <- bind_cols(enframe(ContrasteVP50), enframe(ContrasteVelevacion)) %>%
rename(CVP50 = `value...2`, CElevacion = `value...4`)
## New names:
## • `name` -> `name...1`
## • `value` -> `value...2`
## • `name` -> `name...3`
## • `value` -> `value...4`
ggplot(Contraste, aes(CElevacion, CVP50)) + geom_point() +
geom_smooth(method='lm') + theme_classic() +
ggtitle("Relación entre contrastes de P50 de HbA y de elevación")
## `geom_smooth()` using formula = 'y ~ x'
#Veamos ahora cómo se distribuyen los caracteres sobre la filogenia
#primero definimos el nombre de los datos
P50_2<-setNames(P50_2[,1],rownames(P50_2))
elevacion_2<-setNames(elevacion_2[,1],rownames(elevacion_2))

#y luego obtenemos su distribución sobre la filogenia propuesta
#para el valor de P50
Dist_P50<-contMap(treeR,P50_2,fsize=c(0.5,1),outline=FALSE)
#modificamos la leyenda del gráfico
plot(Dist_P50,fsize=c(0.5,1),outline=FALSE,lwd=c(3,7),leg.txt="P50")
#para las altitudes máximas registradas para cada taxón
Dist_elevacion<-contMap(treeR,elevacion_2,fsize=c(0.5,1),outline=FALSE)
#nuevamente, modificamos la leyenda del gráfico
plot(Dist_elevacion,fsize=c(0.5,1),outline=FALSE,lwd=c(3,7),leg.txt="Elevacion")
```

Pregunta 4

¿Según los resultados obtenidos, la historia filogenética compartida de las especies parece haber influido en la evolución de P50? ¿La inercia filogenética (o tendencia en la historia del rasgo P50) parece haber impedido la adaptación en algunos pares de spp?

Práctico 3

Procesos de diversificación

Darwin pudo haber sido el primero en describir la radiación adaptativa cuando, contemplando la variedad de pinzones que ahora llevan su nombre, comentó: "*Al ver esta diversidad de estructura en un grupo pequeño e íntimamente relacionado grupo de aves, uno realmente podría imaginar que partiendo desde una escasez original de aves en este archipiélago, una especie ha sido tomada y modificada en diferentes fines*" (Darwin, 1845). Desde la época de Darwin, los naturalistas y biólogos evolutivos han estado fascinados por la extraordinaria diversidad de ecología, morfología, comportamiento y riqueza de especies de algunos clados, pero ahora ha resurgido el interés por la radiación adaptativa.

La explosión de las filogenias moleculares en las últimas tres décadas **(!)** han contado las historias detrás de la diversificación de innumerables clados y ha proporcionado la materia prima para un renacimiento de los estudios de radiación adaptativa (1). Las filogenias moleculares han ofrecido descubrimientos sorprendentes sobre la historia y la magnitud de muchas radiaciones adaptativas, como las vángidas de Madagascar (Vangidae), las córvidos de Australia (Corvidae), los cíclidos del lago Victoria (Cichlidae), las lobelias de Hawaii (Lobelioideae) etc. En cada uno de estos casos, se pensaba que la gran diversidad ecológica y morfológica de un grupo era el resultado de eventos de colonización independientes de múltiples linajes ancestrales adaptados de manera diferente. En cambio, nuevas filogenias moleculares revelaron que la gran diversidad en estos grupos es el resultado de la evolución *in situ*, es decir, una radiación adaptativa.

Con filogenias calibradas en el tiempo, uno puede plantearse si la ocurrencia de radiación está correlacionada con eventos históricos (por ejemplo, extinciones masivas, cambios en el clima) o si el ritmo de diversificación disminuye con el tiempo, como a menudo se espera de una radiación adaptativa.

Con el estadístico gamma, calculado a partir de la topología podemos describir tendencias macroevolutivas como una primera aproximación a ver las tendencias dentro de un grupo taxonómico. Se puede averiguar si hubo tasa de diversificación constante. Se puede detectar cambios en la tasa de diversificación a lo largo del tiempo: si gamma es negativo ~ ha habido una desaceleración en la tasa de diversificación a lo largo del tiempo y si gamma es positivo ~ esto sugiere que ha habido una aceleración en la tasa de diversificación a lo largo del tiempo. También, si el estadístico gamma es significativamente diferente entre dos clados, entonces esto sugiere que las tasas de diversificación de los dos clados son diferentes.

Actividad Práctica

Objetivos

(i) ver un ejemplo de estrategias o métodos que se pueden aplicar luego de obtener un árbol.
(ii) ver cómo emplear estadísticos simples teniendo en cuenta solo la topología y largo de rama de los árboles.
(iii) más en general, discutir cómo la ecología o historia pueden afectar el éxito de un grupo.

Paso 1

Ir a esta página https://birdsoftheworld.org/bow/species
Cada subgrupo elegirá **una** familia de aves y trataremos de ver los patrones de especiación dentro de cada linaje.

A priori ¿conoce algún grupo "especioso"?
¿Conoce algún grupo con una aparición repentina de especies?
Para ver “estos patrones” entre linajes ¿por qué es importante comparar grupos o linajes equivalentes?

Paso 2

Ver esta otra fuente adicional de datos de Aves (la usaremos luego)
https://www.worldbirdnames.org/new/classification/family-index-2/

¿Qué datos podríamos extraer de aquí?

Ver tabla Excel online o Google spreadsheet con las familias disponibles acá
https://tinyurl.com/3jvx6afa

Teniendo en cuenta los datos de este sitio: https://birdsoftheworld.org:

- ➔ Anotar la distribución geográfica de la familia elegida.
- ➔ Anotar el número de especies y géneros.

Paso 3 en R

Cargar el árbol de la familia elegida.
Hacer plot de los linajes a través del tiempo (ltt)
Anotar estadístico gamma y su *p-valor* asociado.
Simular 20 LTTs (para árboles de tamaño similar) y hacer plot del ltt 'real' junto a los 20 simulados.

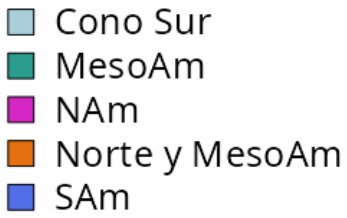
Ahora para la familia que eligieron:

Anotar el estimado de gamma.
Anotar el tiempo de origen de cada Familia.
Ajustar la tasa de especiación y extinción para este árbol. ¿Qué modelo es favorecido *bd* o *yule* (puro nacimiento)? Anotar **"b"** (especiación) y **"d"** (extinción).

Paso 4 …seguimos en R

Ahora nos adentramos en una familia en concreto…
Cargar el árbol de "`Trochilidae.tree`"
Cargar los datos de la distribución "`South.tsv`"
Ver ambos datos, árbol & datos.
¿Teníamos datos de distribución faltantes? ¿Cómo recuperamos esa data?
¿Cómo puede influir la región geográfica en la especiación?
¿Cómo puede influir el tipo de hábitat en la especiación?

Primero, hacemos la reconstrucción del área geográfica.




```
MA ~ America Central (tierra firme)
MA,SA ("cono Sur")  ~ CS
NAm ~ Norte America incluye islas Caribe
NA,MA ~ Norte
SAm ~ SAmerica
```

¿Cómo se codificó el estado de carácter “geografía” en estos datos?

Ahora, viendo el plot de LTT podemos contestar lo siguiente,

¿Cuál es la región que acumula más linajes a través del tiempo? Entonces, ¿influyó la región en la diversificación de los picaflores?
¿Qué otros datos "**geográficos**" o de la distribución de estas especies podemos considerar para examinar los patrones macroevolutivos en estas aves? ¿Qué otros **rasgos** de estas especies podemos considerar para examinar los patrones macroevolutivos en estas aves? Veamos pot ejemplo los datos de una especie dentro de este grupo https://www.iucnredlist.org/species/22687097/166914847

Citas

1. Losos, Jonathan B., and D. Luke Mahler. "Adaptive radiation: the interaction of ecological opportunity, adaptation, and speciation." Evolution since Darwin: the first 150 (**2010**): 381-420.
2. Harmon, Luke J., James A. Schulte, Allan Larson, and Jonathan B. Losos. "Tempo and mode of evolutionary radiation in iguanian lizards." Science 301, no. 5635 (**2003**): 961-964.
3. Jetz, Walter, Gavin H. Thomas, Jeffrey B. Joy, Klaas Hartmann, and Arne O. Mooers. "The global diversity of birds in space and time." Nature 491, no. 7424 (2012): 444-448.
4. McGuire, Jimmy A., Christopher C. Witt, J. V. Remsen, Ammon Corl, Daniel L. Rabosky, Douglas L. Altshuler, and Robert Dudley. "Molecular phylogenetics and the diversification of hummingbirds." Current Biology 24, no. 8 (2014): 910-916.

Práctico 4

Deriva genética

Objetivos.
1- El modelo de Hardy-Weinberg sirve y se utiliza como hipótesis nula para estudiar el grado de alejamiento del supuesto.
2- El tamaño poblacional influye en la probabilidad de fijación o extinción de los distintos alelos.
3- El efecto a largo plazo de la deriva es la pérdida de la heterocigosidad y la mutación revierte ese proceso.

Parte I. Ley de Hardy-Weinberg

El modelo de Hardy-Weinberg describe el comportamiento de las frecuencias génicas de una generación a la otra en poblaciones ideales. Se dice que una población está en equilibrio de Hardy-Weinberg cuando sus frecuencias alélicas y genotípicas se ajustan a las predicciones del modelo.
Bajo los supuestos del modelo, siendo pi la frecuencia del alelo Ai, las predicciones de frecuencias genotípicas serían las siguientes:

$$E(A_i\ A_i) = p_i^2 \qquad E(A_i\ A_j) = 2p_ip_j$$

Las predicciones de Heterocigosidad, para múltiples alelos, estarán dadas por la siguiente ecuación:

$$E(H) = 1 - \Sigma p_i^2$$

Siendo "$\Sigma\ p_i^2$" la suma de las frecuencias esperada de homocigotos, o sea la Homocigocidad.

La Heterocigosis (medida de la variabilidad en la población) aumenta a medida que aumenta el número de alelos en la población.

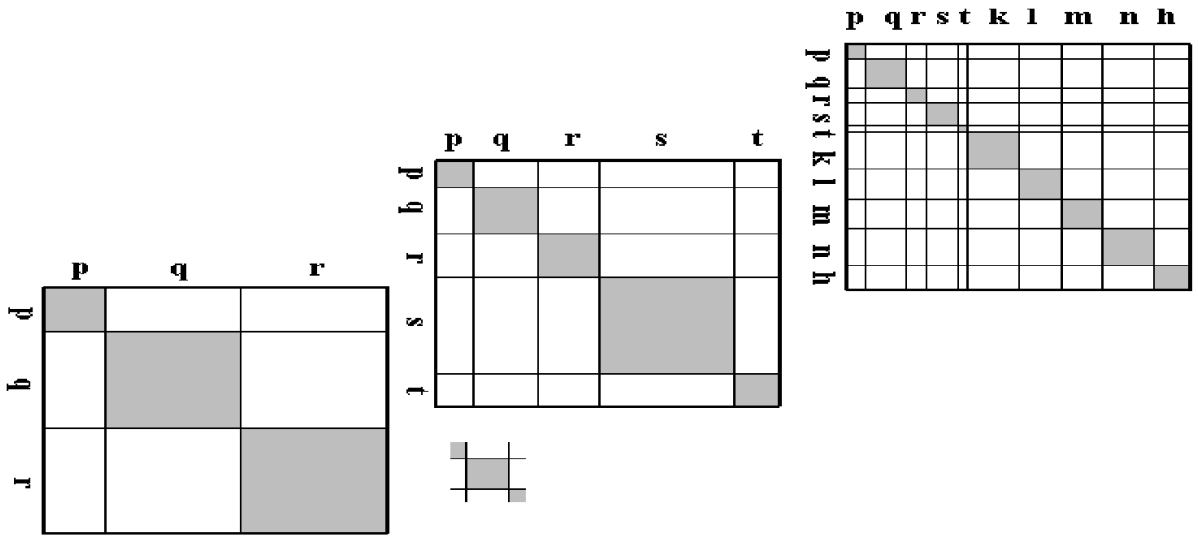
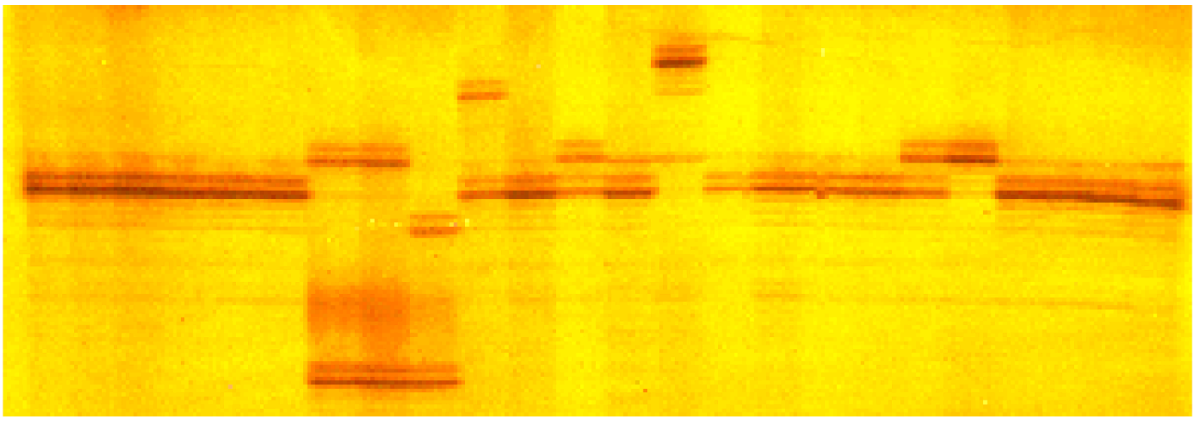


Figura1 El área sombreada representa la fracción de homocigotos esperados bajo equilibrio Hardy-Weinberg.

Caso de estudio- Se estudia la variación genética de una población de *Ctenomys rionegrensis*. A partir del mismo, calcule las frecuencias alélicas y genotípicas observadas globales. Calcule la heterocigosis observada y la esperada a partir del modelo de Hardy-Weinberg considerando a toda la muestra como parte de una única población. Interprete los resultados.



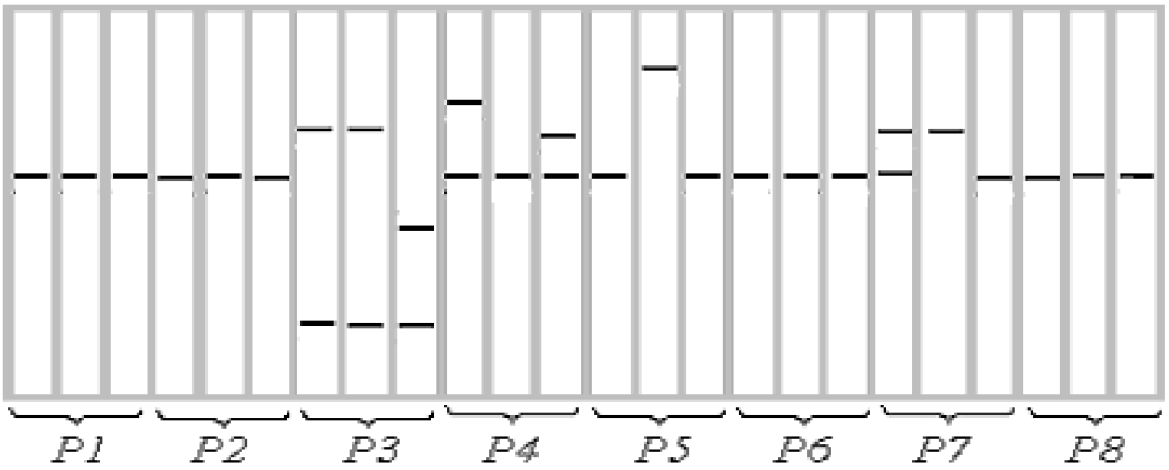


Figura2 Gel de la poliacrilamida mostrando los resultados de la amplificación de microsatélites para tres individuos de cada una de ocho subpoblaciones de Ctenomys rionegrensis (arriba), y su representación esquemática (abajo).

La resolución de este problema se encuentra disponible en los materiales del práctico.

Parte II. Deriva genética

Cuando de los supuestos del modelo de Hardy-Weinberg se abandona la condición de que las poblaciones tienen tamaño infinito, se puede observar que de una generación a otra se producen cambios aleatorios en las frecuencias génicas. Esto es efecto directo de que los alelos que forman una nueva generación son un muestreo al azar de los alelos presentes en la generación parental. En este caso, el sentido de la evolución no se puede predecir, ya que está causada por los efectos de muestreo de generación en generación, un proceso llamado deriva genética.

El modelo poblacional que sólo difiere del de Hardy-Weinberg en poseer un tamaño poblacional finito es el modelo Fisher-Wright, que modela la probabilidad de un alelo de tener determinada frecuencia en la generación siguiente, únicamente a partir de la frecuencia inicial del alelo y el tamaño poblacional, realizando muestreos al azar con reposición.

Para los siguientes ejercicios se utilizará el programa AlleleA1 (https://faculty.washington.edu/herronjc/a1/) que permite simular variaciones en la frecuencia génica en función del tiempo, cambiando parámetros como la frecuencia alélica inicial y el tamaño poblacional, entre otros.

Las simulaciones se pueden hacer para uno o varios loci no ligados cada uno con dos alelos, lo que también se puede interpretar como varios ensayos sucesivos independientes para un solo locus con dos alelos (*Genetic Drift → Number of finite populations to simulate*). En cualquiera de los casos, el programa grafica la frecuencia de uno de los dos alelos del locus, y el otro alelo será el complemento de esa. El botón "*Run Again*" en la ventana gráfica de la derecha permite realizar nuevas simulaciones.

Ejercicio 1. Efecto del tamaño poblacional

¿Cómo crees que es el efecto de la deriva genética según el tamaño poblacional?
Para comparar el efecto de diferentes tamaños poblacionales, realizaremos simulaciones manteniendo una frecuencia inicial de 0.5, para 5 loci, durante 500 generaciones (ventana de la izquierda), y dos tamaños poblacionales diferentes:

Simulación A: N = 500 (gran tamaño poblacional)
Simulación B: N = 100 (pequeño tamaño poblacional)

Anotar por lo menos para una simulación de cada tipo las frecuencias finales de los alelos ¿Cuántos alelos se fijan o eliminan? ¿en cuántas generaciones? Comparar y discutir tus resultados con los de tus compañeros. ¿Son estos resultados coherentes con tus predicciones?

Ejercicio 2. Efecto de la frecuencia alélica inicial

Reflexione acerca de con qué frecuencia inicial se encontrará una mutación que recién surge en un población. ¿Cuál será su probabilidad de fijarse?

Corre algunas veces la siguiente simulación durante 500 generaciones:

Simulación C: N = 500, y una baja frecuencia alélica inicial: 0,1

Nuevamente registre para al menos una corrida, cuántos alelos se fijan o eliminan y en cuántas generaciones. Comparar los resultados obtenidos en esta última simulación con los de la simulación A. ¿Son estos resultados coherentes con sus predicciones?

Ejercicio 3. Efecto a largo plazo sobre la heterocigosis

Realizar una nueva simulación con N=500, por 1000 generaciones, y frecuencia inicial de 0,5, y calcula la heterocigosis promedio mediante la ecuación básica $H = 2pq$ en $t = 0$, 500 y 1000. ¿Existe alguna tendencia en los resultados?

Parte III. Deriva genética y mutación

Ahora introduciremos una fuente de variación abandonando otro de los supuestos de los modelos de Hardy-Weinberg y de Fisher-Wright. Permitiremos el surgimiento de mutaciones neutrales (sin selección) a una tasa μ por alelo por generación. Consideramos que cada variante nueva puede surgir por mutación una única vez.

Ejercicio 4

Reflexione sobre: ¿cuántas mutaciones neutrales, en promedio, se fijarán por deriva en cada generación? ¿Cómo afecta el tamaño de la población a esta tasa de sustitución?
Realizar una simulación inicial con freq. inicial A1 = 0.8, 10 poblaciones, con una tasa de mutación A1 to A2=0,00005, 5000 generaciones y un tamaño poblacional N=100. ¿Qué

observa? Ahora repetimos pero agregamos mutación A2 to A1 = 0,00005 también. ¿Qué observa? Discutir nuevamente que aporta la mutación.

Opcional

Para visualizar una manera de cómo pueden realizarse algunas de estas simulaciones a partir de la distribución binomial, usaremos nuevamente el entorno de programación R con el archivo “Deriva genética.rmd", en donde también podremos cambiar algunos parámetros, pero sabiendo cuál es la base del cálculo.

Práctico 5

Selección Natural y Selección Sexual

Objetivo: Familiarizarse con el concepto de selección natural y algunos métodos para detectar su acción en distintos tipos de datos.

1. Estimación de eficacia y coeficiente de selección a partir de datos de campo

El tucu-tucu de Río Negro presenta fenotipos melánicos, agutís y dorso oscuro, con frecuencia coexistiendo en suelos arenosos. El fenotipo melánico debería ser más fácilmente detectable, y por tanto sufrir mayor mortalidad por depredación. Un estudio de la población de Estancia "La Tabaré", Depto. de Río Negro, –en donde solo coexisten las variedades melánica y agoutí– obtuvo 74 juveniles en una estación reproductiva: algunos en trampas colocadas dentro de las cuevas y otros en bolos regurgitados en torno a un nido de "Lechuza de las vizcacheras" (*Athene cunicularia*) localizado en el mismo campo[1]. En la siguiente tabla, se estima la supervivencia de cada fenotipo descontando los observados en bolos de lechuza del total de juveniles de cada fenotipo[2]:

Fenotipos	Total de juveniles	Sobrevivientes	Eficacia absoluta	Eficacia relativa (w)	Coeficiente de selección (s)	S (%)
Agutís	46	38				
Melánicos	28	19				

Ejercicio

- Calcular la eficacia absoluta (fracción de sobrevivientes), la eficacia relativa w (cociente entre eficacia absoluta de un fenotipo y la correspondiente al fenotipo más apto).
- Calcular el coeficiente de selección (s) para cada fenotipo. ¿Qué significan en términos biológicos los coeficientes de selección calculados?

¿Qué conclusión es posible sacar sobre la mortalidad diferencial de los dos fenotipos por causa de la depredación a partir de estos datos?, ¿cuáles serían sus limitaciones?
¿Qué utilidad tienen estos resultados para entender la eficacia darwiniana general de ambos fenotipos? Razone sobre las posibles limitaciones.
¿Cómo podría explicarse la persistencia del fenotipo melánico? ¿Qué estudios podrían realizarse para avanzar en la comprensión del problema?

Opcional

Ingresa los valores de eficacia relativa obtenidos en el simulador de AllelleA1 utilizado en el práctico anterior (https://faculty.washington.edu/herronjc/a1/).

---

[1] Vasquez Herrera, A. 2003. Posible depredación diferencial sobre individuos agutís y melánicos de *Ctenomys rionegrensis*, reflejada en bolos de *Athene cunicularia*. Informe de Pasantía, Licenciatura en Ciencias Biológicas, Facultad de Ciencias, Universidad de la República, 35 pp.

[2] Datos adicionales: De los 50 bolos estudiados, 26 contenían restos de tucu-tucus. De estos, fue posible determinar el color del pelaje en 17.

2. Selección Sexual

La selección sexual resulta en variación en el éxito reproductivo entre individuos del mismo sexo y típicamente actúa más fuertemente sobre los machos. Puede ser dividida en intrasexual e intersexual y, aunque la evolución de ciertos rasgos de los machos puede ser promovida exclusivamente por uno de los dos componentes, a menudo actúan en forma simultánea. La evidencia empírica sugiere que podrían actuar en direcciones opuestas y el resultado neto reflejaría el equilibrio entre esos dos procesos. Sin embargo, en muchos casos la selección intra e intersexual tienen efectos complementarios, promoviendo la expresión de los mismos rasgos en machos.

El caso de los peces anuales

Austrolebias charrua es una especie de pez anual endémica del sistema de lagunas Patos-Merín. Habitan charcos temporales en cuyos sustratos depositan huevos resistentes a la desecación. La especie presenta un dimorfismo sexual muy marcado: los machos son más grandes, tienen distinta coloración y un patrón de bandas verticales oscuras en los flancos del cuerpo. Un estudio evaluó la selección sexual sobre el tamaño corporal de los machos por medio de la elección de la hembra y la competencia entre machos (Passos et al. 2013[3]) (videos de cortejo y agresión en peces anuales).

Actividad i) Preferencias de las hembras.

Se llevaron a cabo observaciones comportamentales usando el diseño del experimento que se muestra en la figura 1a. Se registró el tiempo que una hembra interactuaba con cada macho, la frecuencia de actividades de cortejo realizada por cada macho y se calculó un índice de apariencia del macho que resume la intensidad del color y el despliegue de la aleta dorsal. Los resultados se ejemplifican en la Tabla 1.

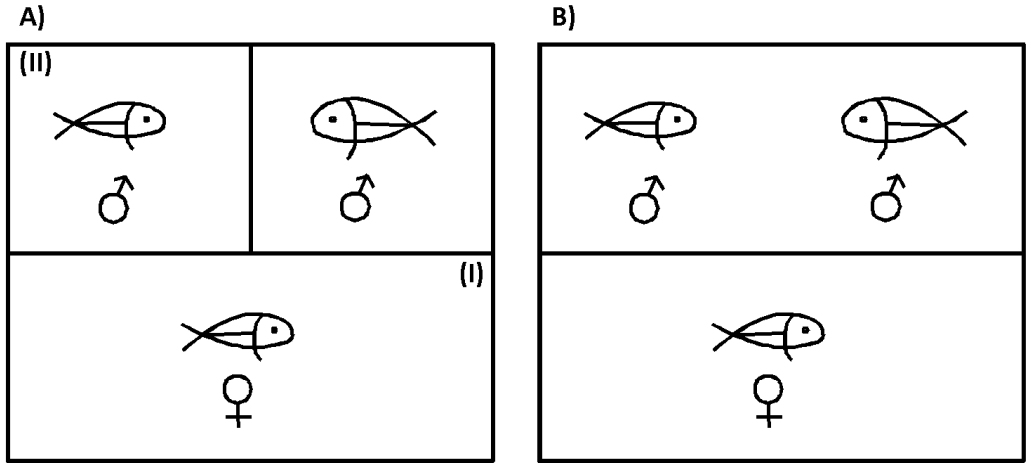


Figura 1- a) Diseño del experimento para evaluar la preferencia de la hembra. La hembra se encuentra sola en un compartimento (i) y puede ver los dos compartimentos de los machos (ii), separados por un tabique que impide que estos se vean entre ellos. b) Diseño del experimento para evaluar competencia entre machos. La hembra se encuentra sola en un compartimento y puede ver los dos machos en el compartimento contiguo.

[3] Passos C, B Tassino, M Loureiro y GG Rosenthal. 2013. Intra- and intersexual selection on male body size in the annual killifish *Austrolebias charrua* . Behavioural Processes 96, 20–26 .

Tabla 1. Tiempo que la hembra interactuó con cada macho (en segundos), la frecuencia de actividades de cortejo realizada por cada macho y el índice de apariencia del macho (2-6). Se muestran 3 casos y se presenta un promedio (última fila) para 30 casos.

Indiv\Variable	**Tamaño** (medido como tiempo)		**Prop. tiempo Cortejo** (medido como prop. tiempo)		**Apariencia Macho**	
	Grande	Chico	Preferido	No Preferido	Preferido	No Preferido
Hembra 1	388	192	0.571	0.909	4	4
Hembra 2	245	145	0.833	0.708	5	5
Hembra 3	284	435	0.613	0.751	5	4
…. .	…. .	…. .	…. .	…. .	…. .	…. .
Promedio	**813***	**425***	**0,683**	**0,641**	**5**	**4**

*denota diferencias significativas (p< 0,05) entre los valores

a) De las características estudiadas, ¿cuál/es condiciona/n la elección de la hembra? Justifique su respuesta.
b) De ser así, ¿cuáles podrían ser las causas de dicha preferencia?

Actividad ii) Competencia entre machos.
Se llevaron a cabo observaciones comportamentales usando el diseño del experimento que se muestra en la Figura 1b (ver video). Se registraron datos de frecuencia de distintos comportamientos agonísticos, el tiempo de resolución del conflicto, la diferencia de tamaño entre los machos (dominante – subordinado) y a partir de ellos se confeccionaron las gráficas de la Figura 2.

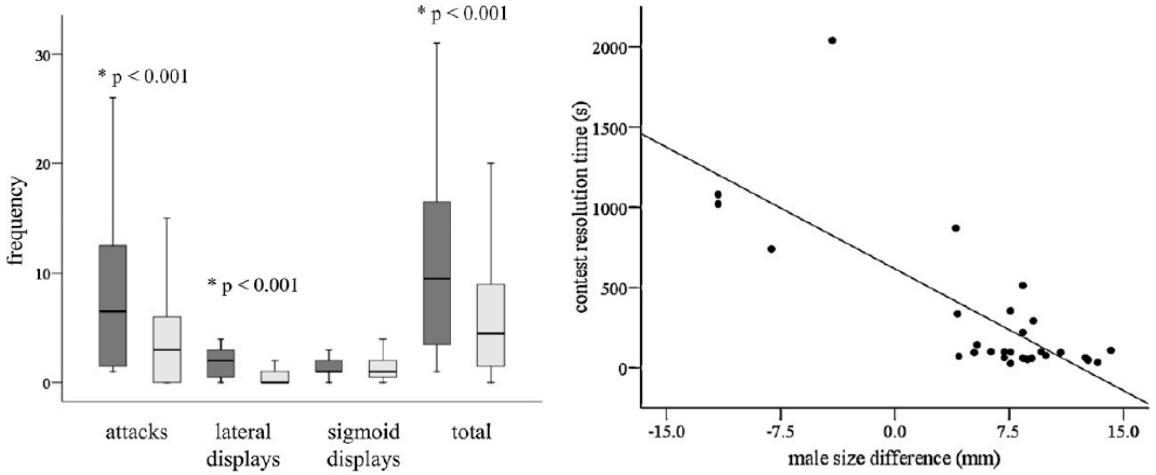


Figura 2. a) Frecuencia de los distintos comportamientos de interacción agonística. En gris oscuro se muestran los datos para los machos más grandes, y en gris claro para los más chicos de cada díada. b) Tiempo de resolución del conflicto (s) en función de la diferencia de tamaño de los machos (mm).

c) ¿Qué diferencias comportamentales se observan entre los machos dominantes y los subordinados?
d) ¿Qué relación hay entre el tiempo de resolución del conflicto y la diferencia de tamaño entre machos?
e) ¿Qué consecuencias fenotípicas tendrían las preferencias de las hembras a nivel poblacional?; y ¿qué otros procesos podrían modificar dichas consecuencias fenotípicas?

Práctico 6

Estructura poblacional y flujo génico

Introducción

Ctenomys rionegrensis es una de las tres especies de este género de roedores subterráneos reconocidas en Uruguay. Su distribución geográfica para nuestro país está restringida a un área de aproximadamente 60 x 50 km al suroeste del departamento de Río Negro (fig. 1).

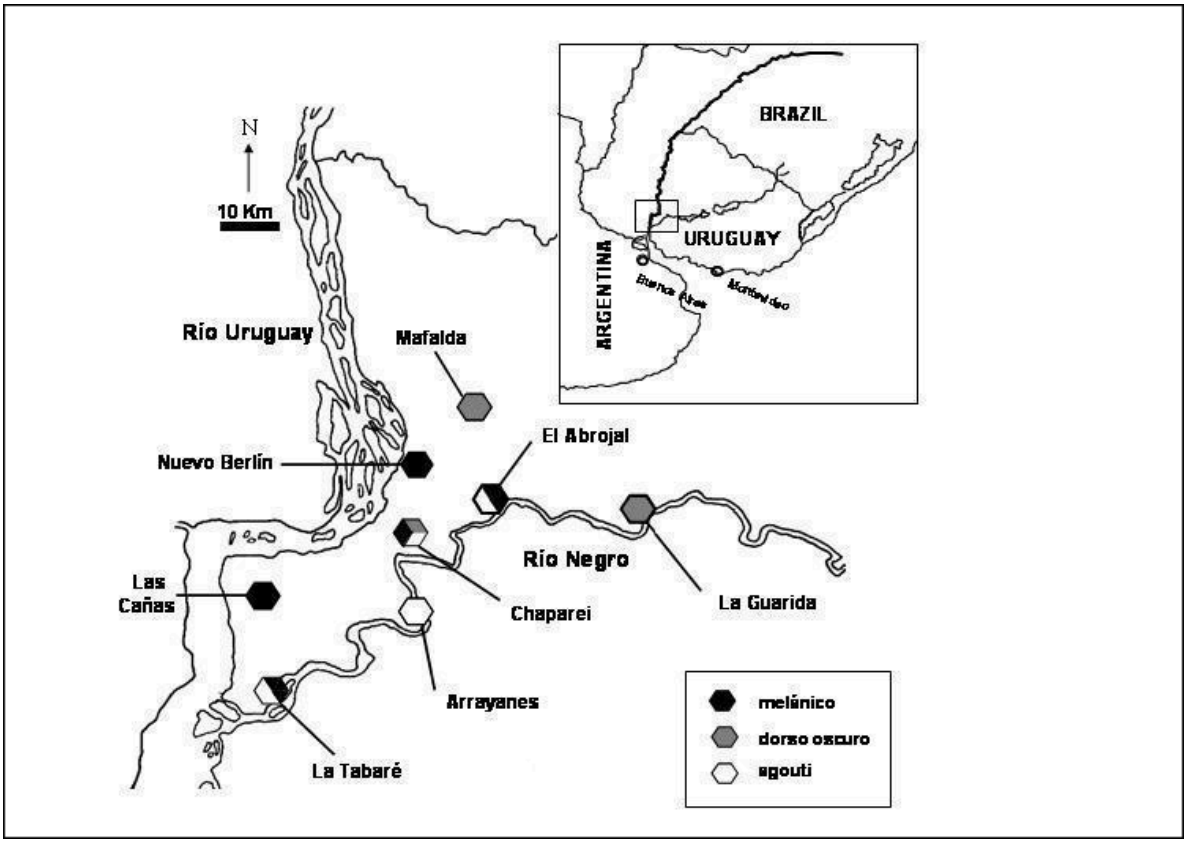


Figura 1. Distribución geográfica de *Ctenomys rionegrensis* en Uruguay. Los sitios que se detallan corresponden a las localidades de muestreo.

Pese a su pequeña distribución, existen tres coloraciones de pelaje marcadamente diferentes: melánico, agutí y dorso oscuro. Mientras que algunas de las poblaciones están enteramente constituidas por individuos que presentan la misma coloración, otras son polimórficas, presentando distintas combinaciones de pelajes.

Esta marcada diferenciación cromática de *C. rionegrensis* es llamativa por varias razones. En primer lugar, por ocurrir en un área geográfica tan reducida. Segundo, no se cumple con la regla general (que incluye a roedores subterráneos) de correspondencia entre color de pelaje y color del sustrato en el que se habita (debida probablemente a presiones selectivas impuestas por la depredación). Todas las poblaciones de esta especie se limitan a la ocupación de suelos arenosos claros, sin diferencias obvias en la vegetación, y de esta forma los individuos melánicos contrastan marcadamente con el sustrato donde viven. A diferencia del comportamiento de otros roedores subterráneos, los tucos emplean más tiempo en la superficie, dedicándose a tareas asociadas al forrajeo y dispersión, incrementando así el riesgo de depredación.

La estructura genética de las poblaciones de roedores subterráneos, como *Ctenomys rionegrensis*, estaría principalmente determinada por:

a) su baja vagilidad, y por lo tanto bajos niveles de flujo génico, y
b) el pequeño tamaño de sus poblaciones, que las hace más susceptibles a los efectos de la deriva genética (operando en mayor grado, fijando o eliminando alelos al azar).

Bajo estas premisas, y dado que hasta el momento no se ha encontrado ninguna posible explicación seleccionista que explique la fijación del pelaje melánico en algunas poblaciones, se ha planteado la hipótesis de una posible fijación del melanismo por deriva genética.

Esta hipótesis prevé:

1) Una reducción de la variación genética en las poblaciones melánicas.
2) Bajos niveles de flujo génico entre las poblaciones, posibilitando que la deriva se sobreponga a los efectos homogeneizadores del primero.

Métodos y Resultados

Para poner a prueba esta hipótesis, se analizaron 11 loci de microsatélites en 150 individuos de *C. rionegrensis* pertenecientes a 8 poblaciones donde se encuentran representados los tres tipos de pelaje, tanto en alopatría como en simpatría (ver mapa).

1) Se calculó la heterocigosis promedio observada Ho y la esperada He para cada población y los estadísticos F de Wright.

Valores promedio	Chaparei	Abrojal	Guarida	Tabaré	Las Cañas	Arrayanes	Mafalda	Nuevo Berlín
# alelos	4	3.18	2.09	3.36	2.27	4.73	3.91	1
Ho	0.54	0.42	0.19	0.43	0.39	0.65	0.54	0
He	0.63	0.46	0.25	0.48	0.39	0.64	0.54	0
F_{IS}								

Actividad 1) Calcular los F_{IS} para cada población.

Los niveles de flujo génico, globales y entre pares de poblaciones fueron estimados de dos formas:

a) Niveles globales: a partir del estadístico F_{ST}, mediante la fórmula $Nm = \frac{\left(\frac{1}{F_{ST}} - 1\right)}{4}$, obteniéndose un *Nm* = 0.45.

b) Estimaciones pareadas: a partir de F_{ST} calculados para cada par de poblaciones.

Para evidenciar si existe un patrón de “aislamiento por distancia”, se construyó el siguiente gráfico con los valores de flujo génico entre pares de poblaciones (obtenidos a partir de las estimaciones pareadas de F_{ST}) en función de la distancia geográfica que las separa (fig. 2). Los valores de F_{ST} fueron estimados para dos marcadores genéticos: microsatélites y el gen del citocromo b del ADN mitocondrial.

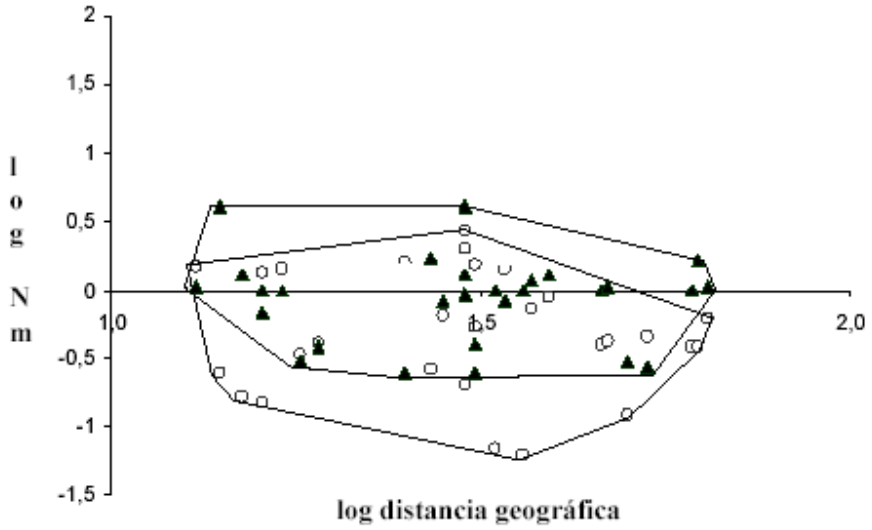


Figura 2. Gráfica del logaritmo del número de migrantes, expresado como Nm, vs. el logaritmo de las distancias geográficas. Los círculos y los triángulos denotan estimaciones basadas en microsatélites y citocromo b, respectivamente.

Además, para el gen del citocromo b se obtuvieron las distintas variantes de las secuencias (haplotipos) y se estudió su frecuencia, su relación con la procedencia de la muestra y la relación entre ellas (fig 3).

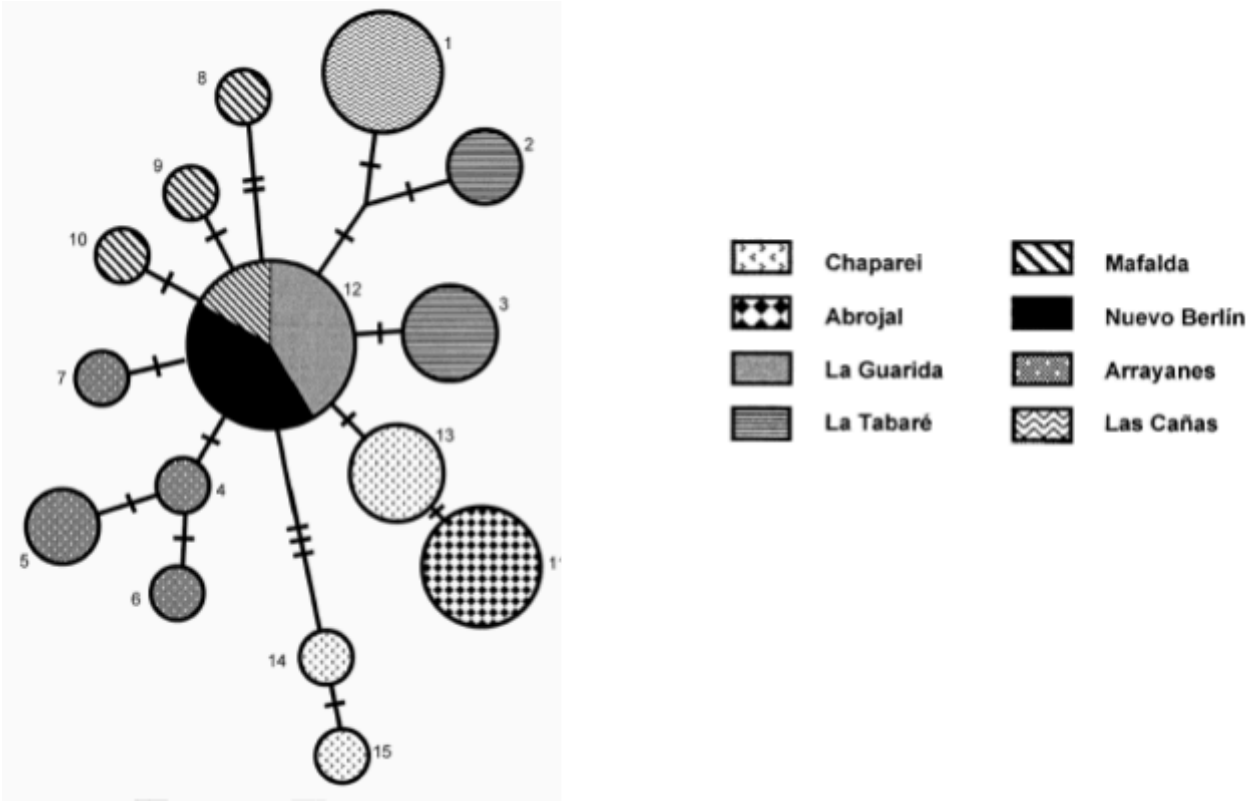


Figura 3. Red de distancias mínimas para los 15 haplotipos de citocromo b encontrados. Cada haplotipo se representa con un círculo, cuya área es proporcional a la frecuencia. Sobre las líneas que conectan haplotipos, los cambios están marcados como rayas transversales. Cada trama representa una población (ver referencia).

Actividad 2) ¿Qué sugiere el gráfico de la Fig. 2 de flujo génico en función de distancias geográficas para pares de poblaciones? (vea también la Fig. 3)

Actividad 3) ¿Las estimaciones de flujo génico obtenidas, corresponden a métodos directos o indirectos? ¿Tiene esto alguna implicancia para nuestros resultados?

Un estudio anterior que considera algunas de las mismas poblaciones y que emplea marcadores alozímicos, propone una estimación de Nm ≈ 6 a 10 y los siguientes valores de F_{IS}:

	Localidad				
	Abrojal	Guarida	Mafalda	Las Cañas	Nuevo Berlín
F_{IS}	0,365	0,577	0,312	0,349	0,242

Actividad 4) Discutir a que pueden atribuirse las discrepancias entre ambas aproximaciones al mismo problema.

Actividad 5) Evaluando la totalidad de los resultados presentados, ¿acepta o rechaza la hipótesis de que el melanismo se fijó por deriva en algunas poblaciones? ¿Qué interpretación plantearías para explicar el patrón de estructura poblacional encontrado?

Bibliografía

Wlasiuk, G., Garza, J. C. y Lessa, E. P. 2003. Genetic and geographic differentiation in the Río Negro Tuco-tuco (*Ctenomys rionegrensis*): inferring the roles of migration and drift from multiple genetic markers. Evolution, 57 (4), pp. 913-926.

Práctico 7

Selección natural: análisis a nivel molecular

Objetivo: Familiarizarse con el concepto de selección natural y algunos métodos para detectar su acción usando datos moleculares.

Introducción

Cuando la selección natural actúa sobre las poblaciones deja huellas que pueden ser reconocidas en el ADN. Para identificar esas huellas se han desarrollado diferentes pruebas aplicables a secuencias nucleotídicas codificantes de proteínas. Una aproximación robusta y sencilla desarrollada por McDonald y Kreitman es considerar que, bajo neutralidad, la relación entre la tasa de cambio nucleotídico sinónimo (dS) y no sinónimo (dN) o de reemplazo aminoacídico será la misma dentro y entre poblaciones. Cualquier desviación sugiere un apartamiento de la neutralidad, incluyendo algún tipo de selección positiva. Si no se cuenta con información poblacional, otra aproximación muy utilizada aunque exigente es considerar que, bajo neutralidad estricta, ambas tasas deberían ser iguales, por lo que dN/dS, también conocido como ω será 1. Si dN supera ampliamente dS, es decir si ω>1, se asume que actuó selección positiva (el caso inverso, ω<1 indicaría selección purificadora). En este práctico aplicaremos ambas aproximaciones.

Ejercicio 1 - **Test de McDonald y Kreitman**

Cuando McDonald y Kreitman en 1991 propusieron su test de neutralidad, lo aplicaron a un conjunto de secuencias de la enzima Alcohol deshidrogenasa (Adh), para tres especies diferentes.

Actividad

La Tabla 1 muestra el resumen de los sitios variables de las secuencias de Adh incluidas en la base de datos.

- Interprete la Tabla 1: ¿qué hay en las filas y las columnas?
- Completar los siguientes cuadros a partir de los datos de la tabla.
- Utilizando los cuadros completados, realizar el test de McDonald y Kreitman (MK) y sacar conclusiones.

	D. simulans vs. D. yakuba	
	Polimorfismos	Sustituciones
Reemplazo		
Sinónimos		

	D. melanogaster vs. D.yakuba	
	Polimorfismos	Sustituciones
Reemplazo		
Sinónimos		

	D. melanogaster vs. D. simulans	
	Polimorfismos	Sustituciones
Reemplazo		
Sinónimos		

Tabla 1. Resumen de los sitios variables de las secuencias de Adh usadas por McDonald y Kreitman (1991) para proponer su test de neutralidad. La primera columna a la izquierda indica la posición del sitio en cuestión, la segunda es la secuencia consenso de referencia; el símbolo – indica un nucleótido idéntico a la secuencia consenso. Se muestran los estados para cada sitio para 12, 6 y 12 individuos de las especies *D. melanogaster*, *D. simulans* y *D. yakuba* respectivamente. Sitios subrayados indican individuos heterocigotas (portan la variante de consenso y la subrayada).

Ejercicio 2 - **Variación en el ω entre sitios y entre linajes**

En este ejercicio usaremos las secuencias de la Hemaglutinina (HA) del virus de la gripe. Esta proteína es una glucoproteína antigénica que se encuentra en la superficie del virus y es la mayor responsable de la unión del virus a la célula infectada. Esta proteína es muy estudiada en el diseño de vacunas, porque presenta una evolución asimétrica que sugiere una fuerte selección de aquellas variantes que son las que mejor escapan al sistema inmune del hospedero. Podemos ver la estructura 3D en este link. Además, el análisis de los cambios nucleotídicos sinónimos y no sinónimos muestra que muchos residuos aminoacídicos en la HA concentrados en el extremo distal y externo de la proteína (que son aquellos sitios que interactúan con el sistema inmune del hospedero) están siendo seleccionados positivamente (Fig. 1).

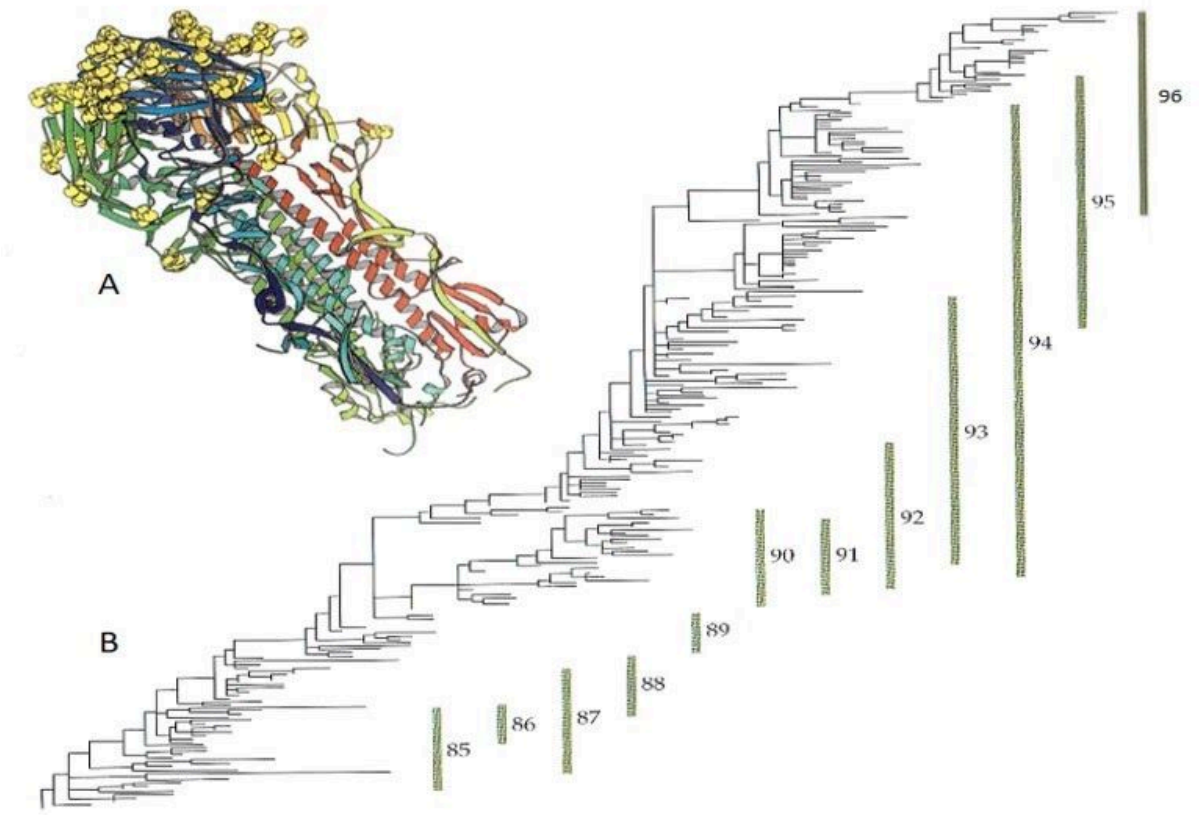


Fig 1. A) Modelo tridimensional de la Hemaglutinina del virus de la gripe, mostrando los sitios aminoacídicos seleccionados positivamente para cambiar. B) Filogenia de cepas del virus aisladas desde 1985 a 1996, basada en el análisis de las secuencias nucleotídicas de ese gen (Tomado de Hillis, 2009).

A continuación, estimaremos la tasa de cambio nucleotídico sinónimo (dS) y no sinónimo (dN) y su relación mediante máxima verosimilitud (ML). Estas estimaciones pueden ser realizadas para cada codón y/o linaje, considerando una filogenia que permita estimar las secuencias nucleótidos en cada uno de los nodos en la filogenia. Dado la gran cantidad de cálculos necesarios para esta tarea, esta estrategia es preferible hacerla en servidores online.
Utilizaremos un servidor llamado Datamonkey (https://www.datamonkey.org/) y usaremos el método SLAC (**S**ingle-**L**ikelihood **A**ncestor **C**ounting). SLAC es una de las tantas formas de evaluar sitios bajo selección. Utiliza una combinación de enfoques de ML y conteos para inferir dN y dS por sitio para un conjunto de secuencias nucleotídicas codificantes de proteínas alineadas y la filogenia que las vincula. SLAC comienza optimizando las longitudes de las ramas y los parámetros de sustitución de nucleótidos bajo un modelo complejo (denominado MG94xREV). Sin embargo, en lugar de usar ML para ajustar los parámetros dN y dS específicos del sitio, SLAC usa ML para inferir la secuencia ancestral más probable en cada nodo de la filogenia. De esta forma se comparan las secuencias nucleotídicas entre nodos adyacentes, y luego cuenta directamente el número total de cambios no sinónimos y sinónimos que se han producido en cada sitio (para eso emplea una versión modificada del método de recuento Suzuki-Gojobori). A su vez, esta aproximación asume que la presión de selección para cada sitio es constante a lo largo de toda la filogenia, algo poco realista.

Actividad

Usamos el archivo fasta SecuenciasHA.fas (en la carpeta de Prácticos). Vamos a ver los resultados en el siguiente link:
https://www.datamonkey.org/slac/6706963c3669ec76d3ab9a50
Interprete los resultados tomando un nivel de significancia de p-value = **0,2.** Ordene los resultados por P[dN/dS > 1] y P[dN/dS < 1].

¿Existe algún sitio y/o linaje seleccionado? ¿Qué sitio/s presenta/n una fuerte evidencia de selección positiva? ¿Qué valores de dN y dS tiene ese sitio? ¿Qué cambios aminoacídicos se registran en ese sitio? ¿Dónde cree que se ubicará ese sitio en la proteína dados los antecedentes planteados.

Ubicar los sitios bajo selección en el esquema “SLAC Site Graph”. En la sección "SLAC Phylogenetic Alignment" visualizar los sitios con más evidencia de selección positiva. Notar los cambios no-sinónimos.

¿Qué similitudes y diferencias encuentra entre SLAC y el test de MK?

Ahora probemos otro enfoque. Podemos usar MEME (**M**ixed **E**ffects **M**odel of **E**volution), que permite estimar selección donde **sólo** algunas de las ramas han experimentado presiones selectivas. Aca permitimos detectar sitios individuales bajo selección episódica diversificadora. Vemos los resultados en el siguiente link:
https://www.datamonkey.org/meme/670696483669ec76d3ab9ab6 .
Ordenar los resultados por p-value.

De forma similar probemos FUBAR (**F**ast, **U**nconstrained**B**ayesian **A**pp**R**oximation). que emplea un enfoque bayesiano para inferir tasas de sustitución no sinónimas (dN) y sinónimas (dS) por sitio para un alineamiento y la filogenia correspondiente. Este método supone que la presión de selección en cada sitio es constante a lo largo de toda la filogenia.
Vemos los resultados en el siguiente link:
https://www.datamonkey.org/fubar/67069cf03669ec76d3ab9bbf .
Ordenar los resultados por BayesFactor[α<β]
Si tenemos en cuenta distintos modelos.. ¿Tenemos resultados congruentes?

Referencias

1. McDonald, J., Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila . Nature 351, 652–654 (1991). https://doi.org/10.1038/351652a0
2. Hillis, DM. (2009). Phylogenetic Progress and Applications of the Tree of Life. En: Evolution since Darwin: The First 150 Years, pp. 421-449. Eds: MA Bell, DJ Futuyma, WF Eanes, JS Levinton. Sinauer Associates, Inc. • Publishers Sunderland, Massachusetts U.S.A.
3. Kosakovsky Pond, SL and Frost, SDW. "Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection." Mol. Biol. Evol. 22, 1208--1222 (2005).
4. Murrell, B et al. "Detecting individual sites subject to episodic diversifying selection." PLoS Genetics 8, e1002764 (2012).

Práctico 8

Patrones de evolución molecular

Objetivos: A partir del análisis del patrón de sustituciones nucleotídicas de una secuencia codificante en un grupo taxonómico particular:

- visualizar patrones generales de evolución molecular
- discutir la validez y el alcance de la idea de "reloj molecular", identificando factores que pueden producir desviaciones aparentes del mismo.

El archivo Primates_datos.meg contiene los 1000 primeros sitios del gen del citocromo b del ADN mitocondrial de 13 especies de primates.

Clasificación de los primates considerados en este práctico

Suborden	Familia	Género	Nombre común	Distribución
Strepsirrhini	Lemuridae	*Lemur* *Microcebus*	Lemur	Madagascar
Platyrrhini	Cebidae	*Cebus* *Saimiri*	Mono capuchino Mono ardilla	Neotrópico
Catarrhini	Cercopithecidae	*Papio* *Macaca*	Babuino Macaco	África
	Hylobatidae	*Hylobates*	Gibón	Asia, Indonesia
	Hominidae	*Pongo* *Gorilla* *Pan* *Homo*	Orangután Gorila Chimpancé Humano	Borneo África África Cosmopolita

Tiempos de divergencia

Datos paleontológicos sugieren los siguientes tiempos de divergencia[4] desde el ancestro común (dados en millones de años desde el presente):

58 Ma. ~ Lemúridos vs. los restantes primates

40 Ma. ~ Platirrinos vs. Catarrinos

15 Ma. ~ Orangután vs restantes homínidos

6 Ma. ~ Gorila vs. Chimpancés y Humanos

[4] En la discusión sobre el reloj molecular y temas relacionados se habla de "divergencia" para referirse al cambio total que ha ocurrido en la evolución de dos especies desde su ancestro común. Este cambio se cuenta, por tanto, a lo largo de dos líneas evolutivas; bajo la hipótesis del reloj molecular la "tasa de divergencia" de un gen o región cualquiera es el doble que la "tasa de evolución".

Actividades

(1) Usando el programa Mega X

Una vez abierta la base de datos “primates_datos.meg” en el programa, realizar las siguientes actividades.

- ¿Por qué será útil indicarle al programa el carácter codificante de la secuencia? ¿y que el origen de la secuencia sea ADN mitocondrial de mamíferos?

- ¿Las secuencias aminoacídicas son más o menos informativas que las secuencias nucleotídicas?

- Obtener una filogenia usando el criterio de Máxima Parsimonia (utilizando las opciones que vienen por defecto). Definir como grupo externo a Lemuridae, reportar el índice de consistencia (en i > summary > CI). ¿Qué información aporta este índice acerca de la filogenia?

- Representar el árbol anterior como filograma (por defecto aparece un cladograma). Reportar si existen diferencias entre grupos en la tasa de evolución y reflexionar las posibles causas que pueden producirlas.

- Obtener una tabla de distancias absolutas pareadas. Escoger en el menú la opción Distances, Compute Pairwise y elegir la opción Model/Method/No. of Differences.

Visualizar las otras opciones.

- Observar la copia de la matriz obtenida anteriormente que se encuentra a continuación. Luego: a) A qué comparación corresponden las distancias resaltadas en gris? b) en la matriz reconocer los recuadros para las 3 comparaciones restantes con que se cuenta con información paleontológica.

		1	2	3	4	5	6	7	8	9	10	11	12
1	*Lemur_catta*	-											
2	*Microcebus_griseorufus*	170	-										
3	*Cebus_albifrons*	245	265	-									
4	*Saimiri_sciureus*	265	267	192	-								
5	*Macaca_mulatta*	263	281	255	272	-							
6	*Papio_hamadryas*	262	278	266	260	135	-						
7	*Hylobates_agilis*	254	263	249	246	203	209	-					
8	*Hylobates_lar*	251	262	258	243	206	215	55	-				
9	*Pongo_pygmaeus*	238	276	248	255	198	207	184	181	-			
10	*Gorilla_gorilla*	239	262	245	240	190	208	163	172	146	-		
11	*Homo_sapiens*	241	270	253	253	199	213	164	170	147	127	-	
12	*Pan_paniscus*	248	275	224	243	188	197	159	170	142	117	112	-
13	*Pan_troglodytes*	239	276	225	236	189	201	164	166	142	120	115	50

(2) Usar el programa Excel

- Abrir el archivo “distancias primates.xls”. Encontrarás el número de diferencias discriminadas entre las posiciones del codón, así como entre transiciones y transversiones,

obtenidas de la forma anterior. Estas distancias pareadas se han graficado para cada uno de los tiempos de divergencia. Observe los rangos de valores para cada una de estas medias, y saque conclusiones de las dos gráficas.

- Pensar, discutir y responder: para el caso de la cantidad de cambios según las posiciones del codón, ¿qué gráfico esperaría obtener según el reloj molecular? ¿Se ajustan en apariencia las gráficas a la idea del reloj molecular? ¿Qué factores pueden dar cuenta de las variaciones observadas?

- De ser posible, estimar el tiempo de divergencia de los gibones y los homínidos.

3) Mencione los problemas asociados a la estimación del tiempo de divergencia entre gibones y homínidos, realizada anteriormente. De encontrarse un pseudogen del citocromo b para estas especies, discuta cómo espera que sea el patrón de sustituciones nucleotídicas en esta secuencia. Bajo neutralidad, ¿es esperable obtener un gen, o región de un gen, con mayor cantidad de cambios no sinónimos que sinónimos?, ¿qué interpretación podría darle a este fenómeno?

Práctico 9

Evolución en familias multigénicas - “Globinas”

Objetivo: Mediante el estudio de los genes de Globinas en primates, familiarizarse con la identificación de familias multigénicas, su problemática y limitaciones.

Una familia génica o familia multigénica es un grupo de loci cromosómicos cuya secuencia de nucleótidos es similar y derivan de una secuencia común ancestral. Puede incluir copias de genes ligeramente diferentes y/o pseudogenes más variables, en uno o varios cromosomas.
Los genes codificantes para las globinas representan un ejemplo clásico de familia multigénica. En general, estas proteínas portan un grupo “heme” y se caracterizan por unirse y transportar oxígeno. Presentan dominios homólogos en varios taxa: a) hemoglobinas tetraméricas de vertebrados (componente proteico principal de los eritrocitos), b) flavohemoglobinas en microorganismos, c) hemoglobinas homodiméricas bacterianas, d) leghemoglobinas en plantas (asociadas al metabolismo del nitrógeno en plantas con rhizomas), e) hemoglobinas de invertebrados, f) mioglobinas monoméricas usualmente encontradas en el músculo animal entre otras.
En primates existen varias clases de globinas dentro de la familia, generalmente designadas con letras griegas (por ejemplo, α, β, γ, δ, ε, ξ y Mioglobina MB). Recientemente se ha identificado una nueva clase, la citoglobina (CYGB). En humanos, el gen que la codifica parece expresarse en todos los tejidos y se localiza en el brazo largo del par cromosómico 17. Se sugiere que su tamaño está conservado en varios mamíferos y es de 190 aminoácidos.

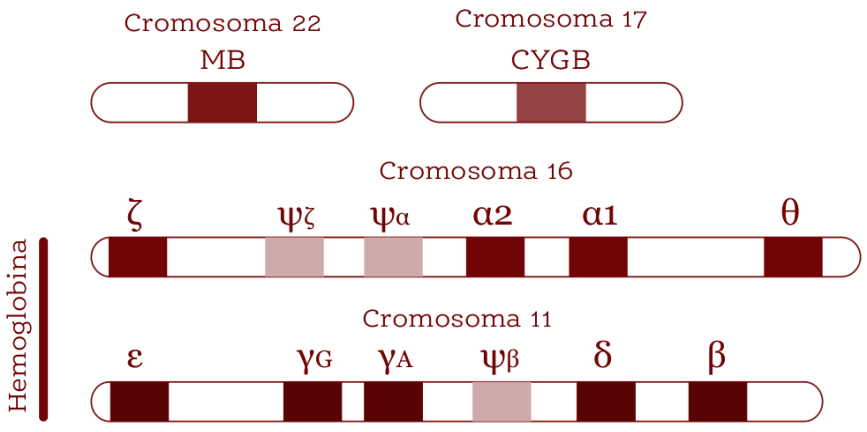


Figura 1 - Localización de los genes codificantes para globinas en humanos.

Conceptos generales

- ➔ ¿Qué procesos podrían estar involucrados en la aparición de las familias multigénicas?
- ➔ ¿Qué ventajas podrían conferir estas familias génicas?

Alineamiento e identificación de globinas

Análisis del alineamiento de secuencias proteicas de globinas, identificación de regiones conservadas
Cargue el alineamiento en el programa MEGA X: Ir a ***Align → Edit/Build Alignment → Retrieve a sequence from a file*** → cargar el archivo "**globinas_aa.fas**" y elegir **Alignment by Muscle**. Ahora nos detenemos unos segundos a ver la estructura de los alineamientos.

- ➔ Identifique los sitios conservados en todas las secuencias y sugiera su posición en la proteína. Justifique.

- ➔ ¿Visualiza algún patrón particular en la secuencia de estas proteínas? Interprete.

- ➔ Selecciones y copie (ctr + c) la región conservada (al menos 50 aa.) en una de las secuencias y haga un blastp (blast de sec. aminoacídicas) contra la base de datos en UniProt (***https://www.uniprot.org/blast***). ¿A qué corresponde dicha región?

- ➔ En la sección Structure podemos apreciar la estructura 3D de esta proteína.

Cargue los datos en el programa MEGA (***File / Open Data /*** **globinas_nt.meg**). Esta base de datos tiene las secuencias codificates (ADNc) reportadas de genes miembros de la familia de las globinas en los siguientes primates: *Homo sapiens, Gorilla gorilla, Pongo abelii, Macaca mulatta, Callithrix jacchus, Papio anubis, Pan troglodytes, Microcebus murinus* y *Otolemur garnettii* (ver Figura 2).

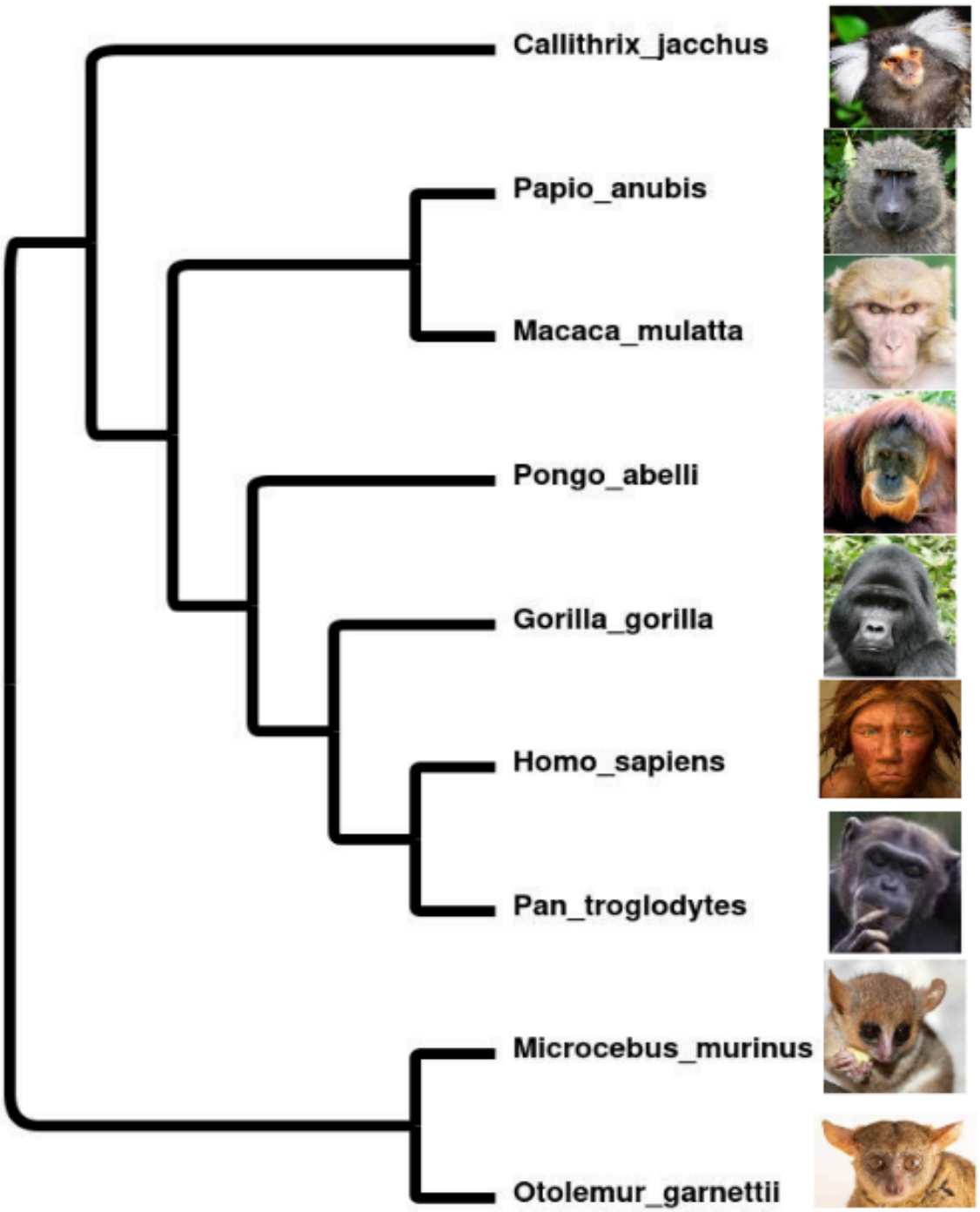


Figura 2 - Cladograma de las especies incluidas los datos

➔ En base a los datos disponibles (ver Fig 1 y 2), ¿cómo piensa usted que ocurrió la evolución de los diferentes tipos de globinas dentro de los primates? Genere un esquema.

Distribución filética de los parálogos

Realice la reconstrucción filogenética en base a los datos que se le brindan (Método: *Neighbor-Joining*, Gaps: *Pairwise deletion,* Bootstrap: *100 pseudoréplicas*). Seleccione Layout: Auto-size Tree.

➢ ¿Son los resultados coincidentes con el esquema que usted generó?
➢ Identifique aquellos nodos que se corresponden con eventos de especiación y/o duplicación.
➢ Analice e interprete algunos casos particulares:
  - clase delta en humanos
  - ausencia de γ-globina en *M. murinus* y *O. garnettii*
  - ausencia de μ-globina en *M. murinus*
  - relación de las citoglobinas con las otras globinas

➔ A grandes rasgos ¿a qué se corresponde la distribución filética de los distintos parálogos de globinas?

➔ Si tenemos que buscar de nuevo un parálogo ¿en cuál cromosoma buscaría la μ-globina en humanos?

➔ ¿Qué información adicional sería útil para establecer el origen de un nuevo gen de globinas?

Estimación de las tasas de evolución en los distintos parálogos

Estimación del parámetro ω (dN/dS) entre clases y dentro de las clases
Los genes ya están agrupados en las clases correspondientes en la base de datos. Calcular las distancias media dentro (ortólogos) y entre clases (parálogos).

➔ ¿Cuál espera sea el resultado dentro y entre clases? ¿Por qué? ¿Qué relación tiene esto con el árbol reconstruido? ¿Cómo piensa que será la estimación de la tasa sinónima y por qué?

En MEGA… Ir a ***Distances*** → ***Compute within group mean*** y luego ***Compute between group mean***, incluyendo sustituciones sinónimas (dS) y no-sinónimas (dN), estimadas con el modelo de Kumar (*Kumar method)*) y Gaps: *Pairwise deletion*.

Distance Estimation

Option		Setting
ANALYSIS		
Scope	→	*Within group average*
ESTIMATE VARIANCE		
Variance Estimation Method	→	*None*
No. of Bootstrap Replications	→	Not Applicable
SUBSTITUTION MODEL		
Substitutions Type	→	*Syn-Nonsynonymous*
Genetic Code Table	→	*Standard*
Model/Method	→	*Kumar method (Kimura 2-para)*
Fixed Transition/Transversion Ratio	→	Not Applicable
Substitutions to Include	→	s: Synonymous only
RATES AND PATTERNS		
Rates among Sites	→	*Uniform Rates*
Gamma Parameter	→	Not Applicable
Pattern among Lineages	→	*Same (Homogeneous)*
DATA SUBSET TO USE		
Gaps/Missing Data Treatment	→	*Pairwise deletion*
Site Coverage Cutoff (%)	→	Not Applicable

- ➔ Abrir el archivo *Cálculo dn-ds familias multigénicas.xlsx*
- ➔ A partir de los valores de distancia media sinónima y no-sinónima calcule la relación entre ellas (≈ dN/dS) dentro de cada parálogo (entre ortólogos) y entre parálogos, y tome nota.
- ➔ Interprete ambos resultados, relacione los mismos con procesos evolutivos dados en clase

Práctico 10

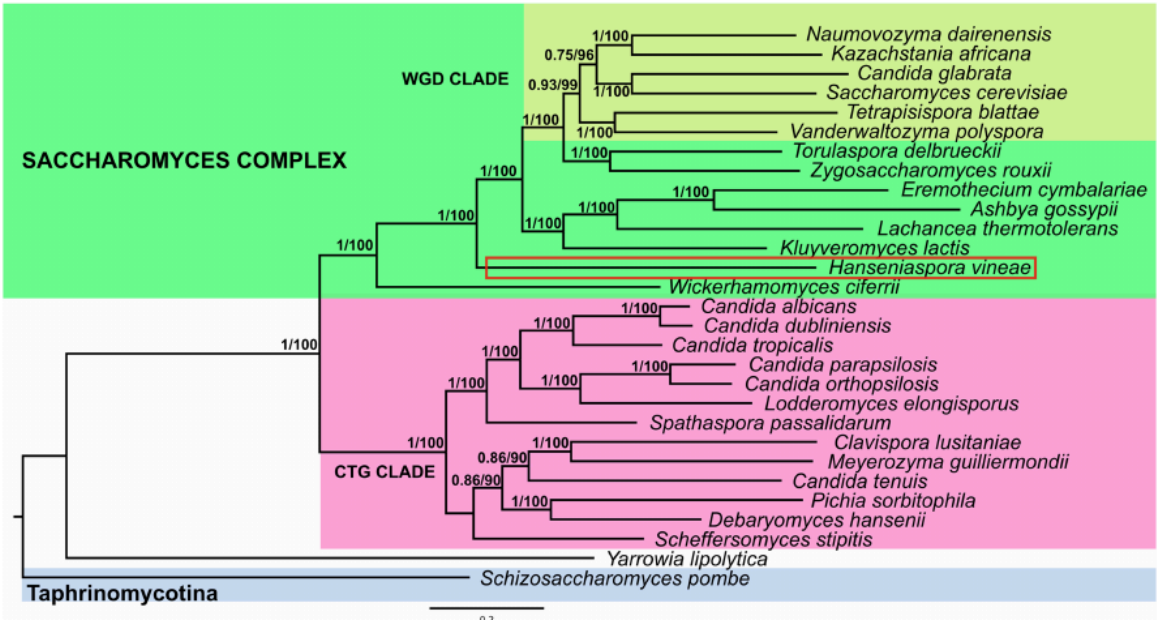
Genómica Comparada

Introducción

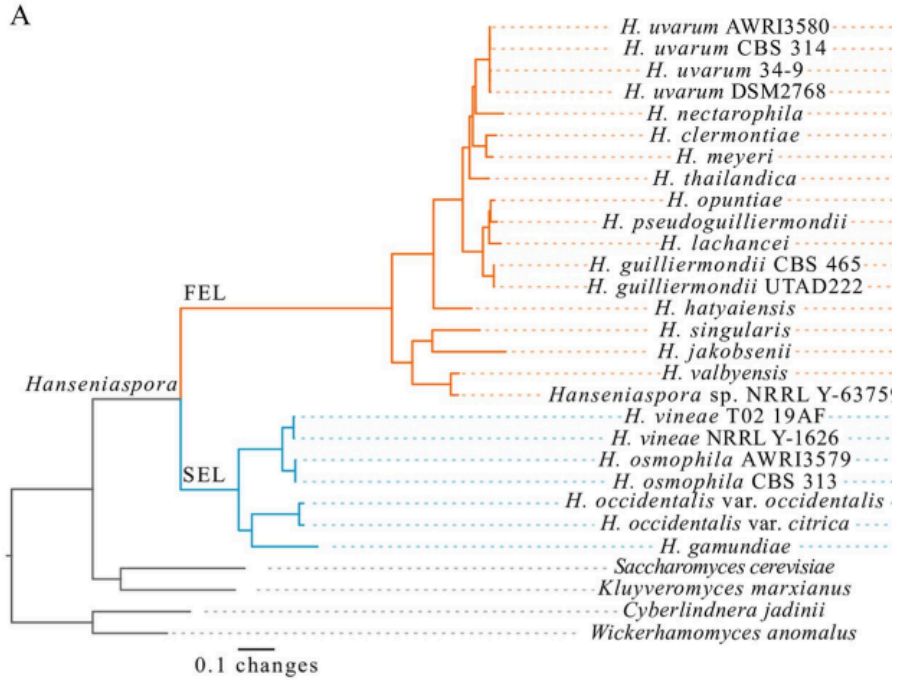
Saccharomyces cerevisiae ('levadura de fermentador' o 'levadura de horneado') es una especie de levadura (hongo unicelular). La especie ha sido fundamental en la elaboración del vino y cerveza y en el horneado desde la antigüedad. *Saccharomyces* fue el primer eucariota cuyo genoma fue completamente secuenciado (¡1996!).

Otro género cercano a *Saccharomyces* es *Hanseniaspora*. Estas levaduras apiculadas se encuentran no sólo en las uvas, sino también en muchas otras frutas. Las levaduras apiculadas representan aproximadamente el 70% de la microbiota asociada a la uva. *Hanseniaspora vineae* está emergiendo como una especie prometedora para la producción de vino de calidad. Los vinos producidos por una mezcla de estas levaduras *H. vineae* con *S. cerevisiae* exhiben consistentemente sabores frutales más intensos y más complejos que los vinos producidos sólo por *S. cerevisiae*. Las especies de *Hanseniaspora* también son participantes clave en la fermentación de una gran variedad de otros productos alimenticios que van desde el chocolate hasta el tequila o la sidra de manzana.

El artículo en el que nos basamos (Giorello *et al.* 2018) detectó que *Hanseniaspora vineae* NO forma parte de una ronda de duplicación que abarca un linaje de 6 especies dentro de la radiación de estas levaduras, donde se encuentra *Saccharomyces* (ver figura). Este trabajo encontró numerosos genes relevantes (ej., ejemplo aminotransferasas y descarboxilasas) que son resultado de duplicaciones génicas.



Más recientemente se encontró que dentro del género *Hanseniaspora*, *H. vinae* es parte de un linaje de evolución lenta (slow evolving lineage, SEL), de aparición más reciente (ver figura, Steenwyk et al, 2019).



Objetivo

Ahora tenemos como objetivo entender la evolución de las levaduras apiculadas y ver cómo ha surgido la diversidad funcional (o distintas funciones) en dos clados dentro de estas levaduras. Queremos también entender si el clado de evolución rápida presentó un patrón evolutivo distinto al de evolución lenta. Para esto analizaremos el genoma completo de 10 levaduras e identificamos qué clados sufrieron más duplicaciones e identificar que funciones cumplen aquellos genes duplicados.

Metodología

Tomaremos algunas de las especies utilizadas en el trabajo de Steenwyk et al. y realizaremos un análisis genómico y de anotación funcional, análisis de ortología sumado a análisis de pérdida y ganancia de genes. Se considerarán *Hanseniaspora guilliermondii, H. occidentalis, H. osmophila, H. valbyensis, H. vineae, Cyberlindnera jadinii, Kazachstania servazzii, Saccharomyces cerevisiae, Wickerhamomyces anomalus* (ver diapositivas)

Parte 1. Acceso a datos genómicos.

a) Localizar uno de los dos genomas de levadura de uva de vino Tannat de Uruguay (ir a https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=29832). Ver el número de genomas depositados para este género.

b) En Uniprot (https://www.uniprot.org/) buscar las proteínas de interés que se encuentran en la siguiente tabla, restringiendo la búsqueda a *Saccharomyces*. Por cada gen, van a “Sequence” y anotar el código/acceso de EnsembFungil.

GEN	Acceso de Ensembl	Nombre común o Función
PDC1	*YLR044C*	
ARO8		
ARO9		
ARO10		
ATF2		
SLI		

Ejemplo para PDC1:
Sección Sequence > Genome annotation databases > EnsemblFungi > YLR044C.

Parte 2. Exploración de sintenia

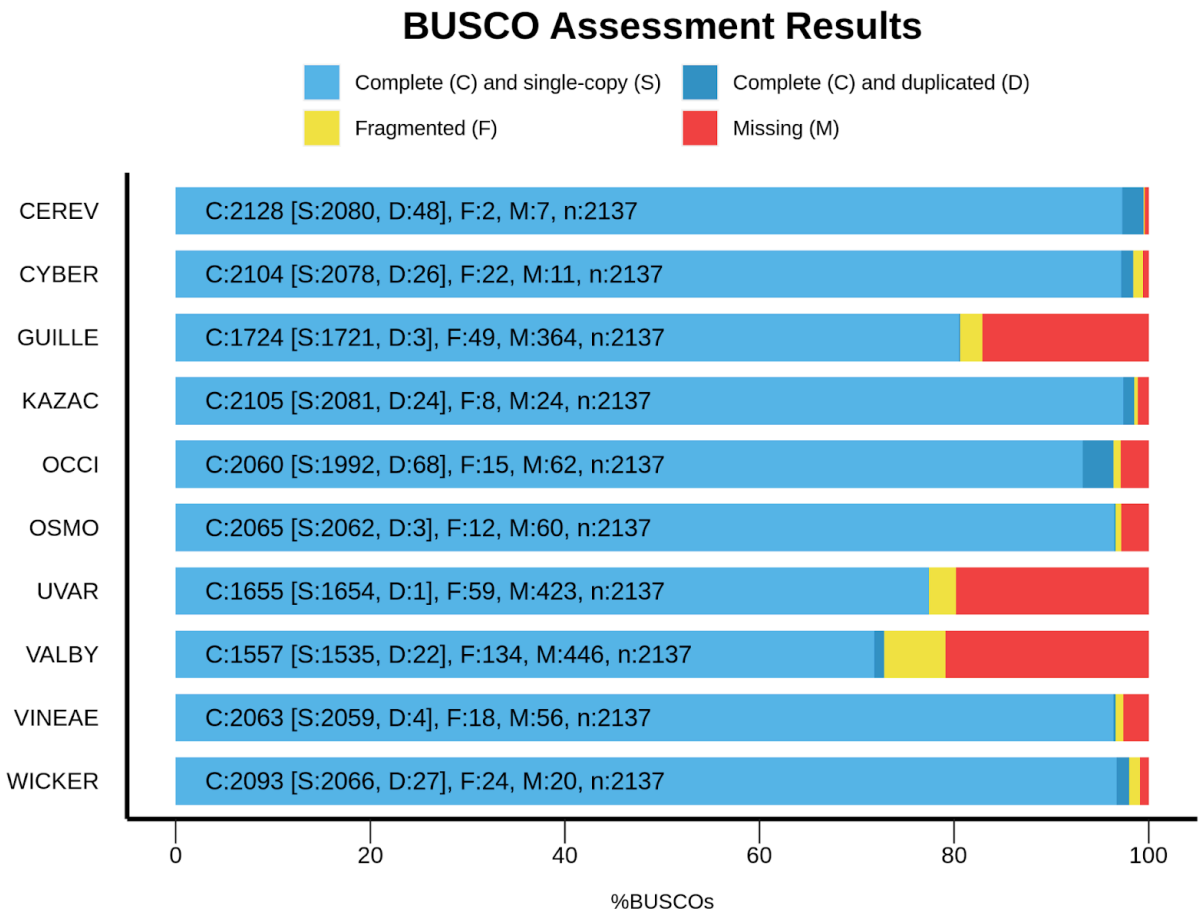
Observar el contexto genómico en "**Genomicus**" [https://www.genomicus.bio.ens.psl.eu/genomicus-fungi-43.01/cgi-bin/search.pl].

En **Genomicus**, buscar el código EnsemblFungi que conseguimos arriba + *S. cerevisiae* S288C.

Identificar qué especies estamos comparando, cuánto se mantiene la sintenia y en que cromosomas se encuentran estos genes. Ya que la visualización es más amigable, identificar duplicaciones génicas en alguno de estos genes (nodos rojos: duplicación; nodos azules: especiación).
¿Aproximadamente, cuántos genomas logramos comparar por este método? ¿Están estos ortólogos en el mismo cromosoma? ¿Qué necesitamos para hacer buenas comparaciones o sacar buenas conclusiones con este enfoque o estrategia?

Parte 3 Calidad del genoma y anotación funcional

Con la herramienta **BUSCO**, vemos cuán completo está cada cada genoma a partir de una base de datos de genes ortólogos de copia única. Si sólo consideramos genes de este tipo, ¿cómo se han originado estos genes? Cuando hacemos "**B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs" (**BUSCO**) logramos cuantificar la “completitud” de un genoma ya que la expectativa es encontrar estos genes en el genoma como "copia única".



Observando los resultados de **BUSCO (a)** ¿Qué genoma se encuentra mejor representado (i.e. más completo)? **(b)** ¿Qué desventaja tiene incluir genomas poco completos?

[Opcional, no lo hacemos!] Ver los resultados de la anotación funcional en *Scerevisiae.xlsx* **(carpeta Parte3 >** tablas 'Scerevisiae.xlsx'***).***

Parte 4. Análisis de ortología y anotación funcional (en R).

Por un lado, **OrthoFinder** estima los ortólogos entre 10 genomas de levaduras. **eggNOG** nos permitió hacer la anotación funcional o conocer qué función cumplen los genes presentes en estos genomas.

Cuando procesamos parte de los resultados logramos más detalle de la historia evolutiva de estas especies. Algunos de los resultados son: un árbol de especies, múltiples árboles de genes, lista con genes ortólogos, lista con duplicaciones génicas y todos los archivos fasta con los ortólogos **(carpeta Parte4)**.

Para esta actividad práctica usaremos, **R** (https://cran.r-project.org/). Los paquetes de **R** (y sus dependencias!) ya han sido instalados: *phytools, tidyverse, formattable*.

Ahora comenzamos a seguir el script ***HansVineae.R***

los números usados debajo -con este formato '(**#1**)'- están asociados en el script de **R**.

Siguiendo los comandos en el script podemos ver el árbol de OrthoFinder. ¿Cómo se construyó? Prestar atención al nombre o etiqueta de los nodos **(#1)**

Teniendo en cuenta el análisis global de ortología nos podemos fijar en algunos resultados que ayudan a describir 'el éxito' o cantidad de ortólogos recuperados.

¿Cuál es el número de genes en los ortogrupos? **(#2)**
¿Número de ortogrupos de copia simple?

¿Qué spp tiene más duplicaciones? Ver plot. **(#3)**
¿A qué clado pertenece esta spp.? ¿Son clado FEL o SEL?

¿Qué especies comparten más ortólogos? Para esto vemos la tabla en R o excel donde se compara cada especie y sus ortólogos compartidos. ¿A qué se debe este patrón? **(#4)**

Tenemos acceso a la estadística de ortólogos por cada genoma. Grafiquemos los siguientes resultados. **(#5)**

- ☐ Nº ortogrupos por spp.
- ☐ Nº genes por spp.

Con la anotación funcional podemos hacernos otras preguntas…
Entre todos los genes anotados vamos a buscar genes de interés en estas levaduras, parte de todo este 'arsenal' de genes asociados a la fermentación.

Nos concentramos en una "alcohol dehydrogenase" mencionada en la Tabla 3 de Giorello et al. ¿Cuántas variantes de SPE1 o SPE2 se pudieron anotar? **(#6)** **¿Qué les permite hacer a las levaduras al tener más cantidad de estos genes?**

Ahora nos fijamos en la lista de duplicaciones **+** anotaciones funcionales

¿Cuántos genes (con anotación) tienen duplicaciones? **(#7)**

En qué categorías COG están presentes estos genes duplicados? **(#8)**
¿Qué función cumplen?

Recordando el resumen del artículo, hay genes señalados como relevantes en estas levaduras. Nos concentramos en un gen en particular…

¿Cuántos genes 'ARO' hay en los Ortólogos? **(#9)**

¿Cuántas duplicaciones hay por especie spp? **(#10)**

Si agrego detalle de qué genes se duplicaron en qué especies tengo otra dimensión de la historia del grupo, podemos hacer otras preguntas. De esta forma podemos entender si la diversidad de genes responde a duplicaciones en el ancestro de estos grupos de levaduras o duplicaciones dentro de cada especie.

¿Cuántas duplicaciones hay en todo el género *Hanseniaspora*? Para esto buscamos antes qué nodo es el ancestro común más cercano de este grupo (recordar paso **#1**) **(#11)**

Podemos buscar qué función (COG) cumplen estos genes duplicados.

Finalmente, ahora podemos visualizar las duplicaciones de algunos genes para poner en contexto cuáles son ortólogos y parálogos.

Visualice las duplicaciones para algunos genes como los ARO, PDC1, descarboxilasas, GTPasas o aminotransferasas **(#16)** ¿Qué podemos intuir acerca de la relevancia de la función en algunos casos de duplicaciones?

“Discusión”

¿Qué genomas agregarían para entender aún más la diversidad funcional de estas levaduras?

¿Qué genes o grupo de genes se podrían explorar con más detalle?

¿Qué potencial tiene mezclar en la producción de vinos u otros fermentos levaduras apiculadas (clados FEL & SEL) y otras levaduras?

MINI GLOSARIO

Anotación funcional: el proceso de adjuntar información biológica a secuencias de genes o proteínas.

Aminotransferasas: enzimas que catalizan una reacción de transaminación entre un aminoácido y un α-cetoácido. Son importantes en la síntesis de aminoácidos, que forman las proteínas.

Categorías COG: en la base de datos "Clusters of Orthologous Groups" (COG), cada COG incluye proteínas que se cree que son ortólogas. El propósito de la base de datos COG es servir como plataforma para la anotación funcional de genomas recién secuenciados y para estudios sobre la evolución del genoma. Para facilitar los estudios funcionales, los COG se clasificaron en 17 categorías funcionales amplias.

Descarboxilasas: son liasas de carbono-carbono que agregan o eliminan un grupo carboxilo de los compuestos orgánicos. Estas enzimas catalizan la descarboxilación de aminoácidos, beta-cetoácidos y alfa-cetoácidos.

Deshidrogenasas: una enzima perteneciente al grupo de las oxidorreductasas que oxida

un sustrato al reducir un aceptor de electrones, generalmente NAD+/NADP+[1] o una coenzima flavina como FAD o FMN. Como todos los catalizadores, catalizan reacciones inversas y directas, y en algunos casos esto tiene un significado fisiológico: por ejemplo, la alcohol deshidrogenasa cataliza la oxidación de etanol a acetaldehído en animales, pero en la levadura cataliza la producción de etanol a partir de acetaldehído.

Genes de copia única (o ortólogos copia única): marcadores universales presentes sólo una vez en cada genoma (para un grupo dado, ejemplo vertebrados o fungi).

GTPasas: una superfamilia de proteínas que regulan muchos procesos celulares, como la señalización celular, el transporte vesicular y la regulación de la forma y motilidad celular.

Ortogrupo: es el conjunto de genes de múltiples especies que descienden de un solo gen en el último ancestro común (LCA) de ese conjunto de especies.

Ortólogo: el “mismo” gen en diferentes especies que ha divergido como resultado de la especiación.

Sintenia: la conservación de bloques de orden dentro de dos conjuntos de cromosomas (o genomas) que se comparan entre sí.

BIBLIOGRAFÍA

Giorello, Facundo, et al. "Genomic and transcriptomic basis of Hanseniaspora vineae's impact on flavor diversity and wine quality." *Applied and Environmental Microbiology* 85.1 (2019): e01959-18.

Steenwyk, Jacob L., et al. "Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts." *PLoS Biology* 17.5 (2019): e3000255.