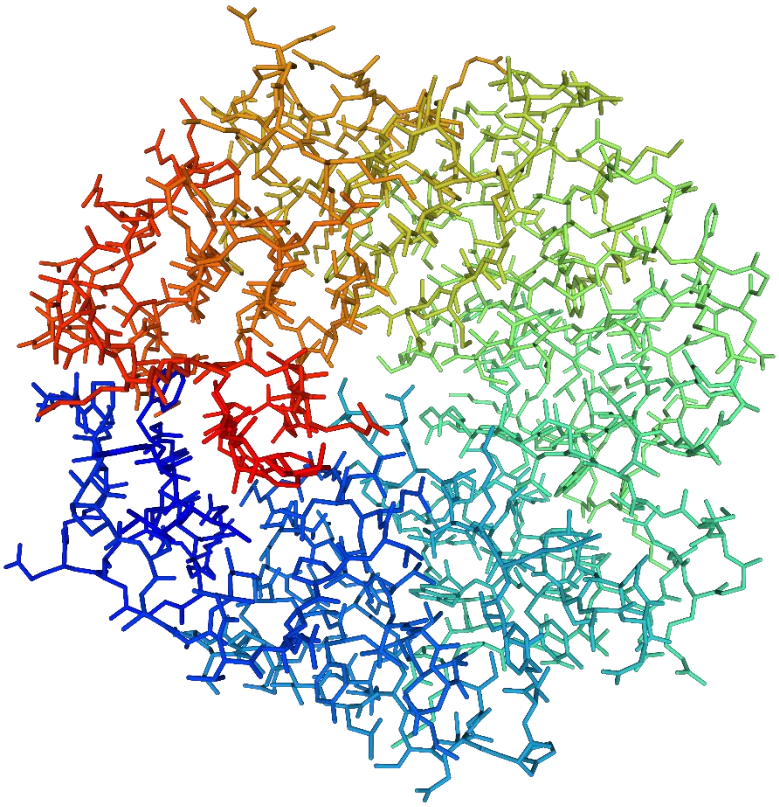
EVOLUCIÓN FILOGENÉTICA DE LA PROTEÍNA HERC2



27 1 2021

CARLOS GARCÍA PERAL

DIPLOMA EXPERTO EN BIOINFORMÁTICA

UNIVERSIDAD DE SALAMANCA

INTRODUCCIÓN

La filogenética molecular es la rama de la ciencia que permite analizar y estimar las relaciones evolutivas a partir de la información genética (genes, proteínas, etc.…) de los organismos. Actualmente se usa cada vez más en los laboratorios de virología para estudiar la transmisión de los virus. Al reconstruir la historia evolutiva de los genomas virales se puede modelar el comportamiento de las poblaciones virales y por ende predecir epidemias futuras.

En el presente estudio nos serviremos de las herramientas de análisis filogenético que proporcionan los paquetes de R (msa y seqinr) para comparan las secuencias homólogas de la proteína HERC2 de 19 especies diferentes que incluyen mamíferos, aves y peces. Dicha proteína pertenece a la familia de las ubiquitinas ligasas E3, caracterizadas por tener en su secuencia aminoacídica un domino HECT y uno o más dominios de tipo RCC1(RCD). Su funcional principal consiste en marcar a otras proteínas con ubiquitinas y entre sus substratos se encuentran las proteínas p53, BRCA2 y determinadas MAP quinasas, por lo que se ha ligado a HERC2 con un rol esencial en numerosos tipos de cáncer. Asimismo, regulan un amplio espectro de funciones moleculares como el neurodesarrollo, la reparación del DNA, el crecimiento celular y la respuesta inmunitaria. Finalmente, cabe destacar que las proteínas de esta familia están altamente conservadas evolutivamente, por lo tanto, detectar las diferencias específicas entre las secuencias de las especies podría servirnos para comprender el mecanismo molecular que permite a las proteínas homólogas desempeñar funciones análogas pero diferentes.

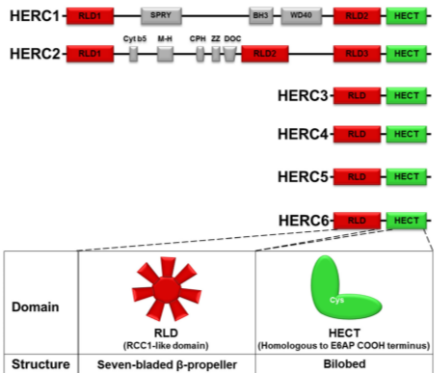


Figura 1. Proteínas de la familia HERC

En primer lugar, se construirá un árbol filogenético que nos permitirá observar las distancias genéticas entre las especies. Posteriormente, examinaremos una región concreta de la proteína HERC2 (el dominio HECT, localizado en el extremo carboxilo terminal), para localizar las posiciones aminoacídicas concretas de la secuencia donde existen discrepancias entre las especies. En este dominio reside la capacidad catalítica de la proteína para unir y marcar diferentes sustratos para su posterior degradación en el proteasoma.

Finalmente se compara específicamente la proteína HERC2 del humano con la proteína de la especie que presente una distancia genética menor. Dicha comparativa nos permitirá identificar las posiciones donde se ha producido una mutación más tardía a lo largo de la línea evolutiva del ser humano.

OBJETIVOS

- Construir un árbol filogenético para establecer las relaciones evolutivas entre los organismos.
- Determinar las principales diferencias en la secuencia aminoacídica en una región del domino HECT
- Comparar la secuencia del humano con la proteína de la especie más cercana

DATOS DE ENTRADA

Para adquirir las secuencias proteicas correspondientes al transcrito mejor anotado del gen HERC2 se empleó el paquete de R biomaRt. Dicho paquete nos permite acceder a las bases de datos disponibles en Ensembl. Posteriormente, se seleccionó la base de datos "ENSEMBL_MART_ENSEMBL", puesto que contiene datasets relacionados con cada especie que queremos estudiar. En nuestro caso se seleccionaron 19 datasets que se corresponden con las siguientes especies: btaurus_gene_ensembl: **toro**, rnorvegicus_gene_ensembl: **rata**, mmusculus_gene_ensembl: **ratón**, mmulatta_gene_ensembl: **macaco**, mgallopavo_gene_ensembl: **pavo salvaje**, scanaria_gene_ensembl: **canario**, ssalar_gene_ensembl: **salmón**, mleucophaeus_gene_ensembl: **gaviota**, preticulata_gene_ensembl: **pez millón**, olatipes_gene_ensembl: **pez japonés**, falbicollis_gene_ensembl: **papamoscas**, mvitellinus_gene_ensembl: **saltarín** y mzebra_gene_ensembl: **pez cebra**.

Una vez seleccionado cada dataset, la función getSequence que proporciona el paquete biomaRt nos ha permitido obtener la información proteica de cada especie. Finalmente, mediante la función exportFASTA los datos de entrada generados con getSequence se guardaron en un fichero FASTA para su posterior análisis.

MÉTODO DE PROCESADO

1) **Alineamiento de secuencias**: para alinear las secuencias de la proteína HERC2 almacenadas en el archivo FASTA se empleó la función msa correspondiente al paquete de bioconductor msa. Con el resultado de dicho alineamiento se elaboraron los siguientes outputs:

- Heatmap: en primera instancia el paquete seqinr permitió calcular una matriz de distancias por pares gracias a la función dist.alignment. Una vez obtenida dicha matriz la función heatmap_rjs correspondiente al paquete RJSplot generó el heatmap final.

- Árbol filogenético: a partir de la matriz de distancias calculada anteriormente se construyó un árbol filogenético empleando los paquetes de CRAN ggdendro y ggplot2. En primer lugar, los datos de distancias se estructuraron jerárquicamente gracias a la función hclust. Posteriormente, la función dendro_data_k permitió separar las especies visualmente en 3 clusters (mamíferos, aves y peces). Finalmente, la función plot_ggdendro mostró el resultado final por pantalla.

2) **Análisis del alineamiento (dominio HECT)**: a partir del anterior alineamiento se extrajo una región aminoacídica correspondiente al domino HECT; posiciones 4540:4610. Dichas regiones se guardaron nuevamente en un archivo FASTA

- Visualización del alineamiento y extracción de la secuencia logo: para visualizar el alineamiento y obtener la secuencia logo se utilizó la función msaPrettyPrint del paquete msa. En este caso los resultados se guardaron en un pdf. Para ello se utilizó la función openPDF del paquete Biobase.

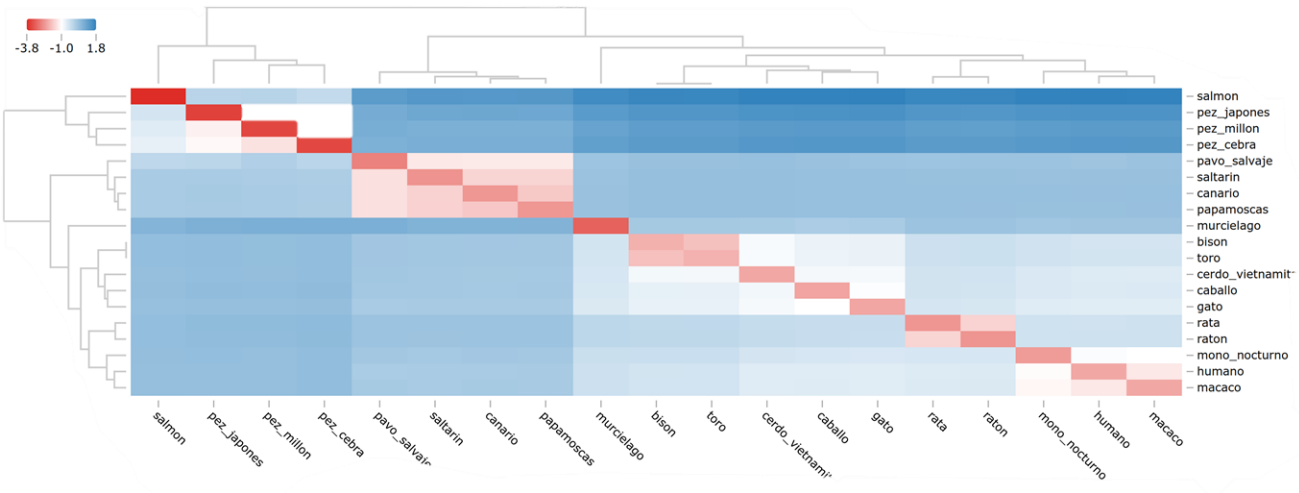
3) **Comparativa entre las secuencias de humano y macaco**: por último, a partir de la información obtenida del árbol filogenético se ha deducido que la proteína HERC2 más parecida a la del humano es la del macaco. De esta manera se han comparado dichas secuencias para determinar en qué posiciones especificas existen discrepancias. Las posiciones se han guardado en un dataframe y se han enviado a un archivo Excel. Además, hay que tener en cuenta que los aminoácidos se pueden dividir en 5 grupos dependiendo de su estructura química, a saber, no polares, polares, cargados positivamente, cargados negativamente y aromáticos neutros. Por ello se ha analizado si dichas diferencias son significativas, es decir, si el cambio se produce por un aminoácido del mismo grupo o no.

- Cambios de aminoácidos: los resultados se han representado gráficamente utilizando el paquete de R ggplot2

DATOS DE SALIDA

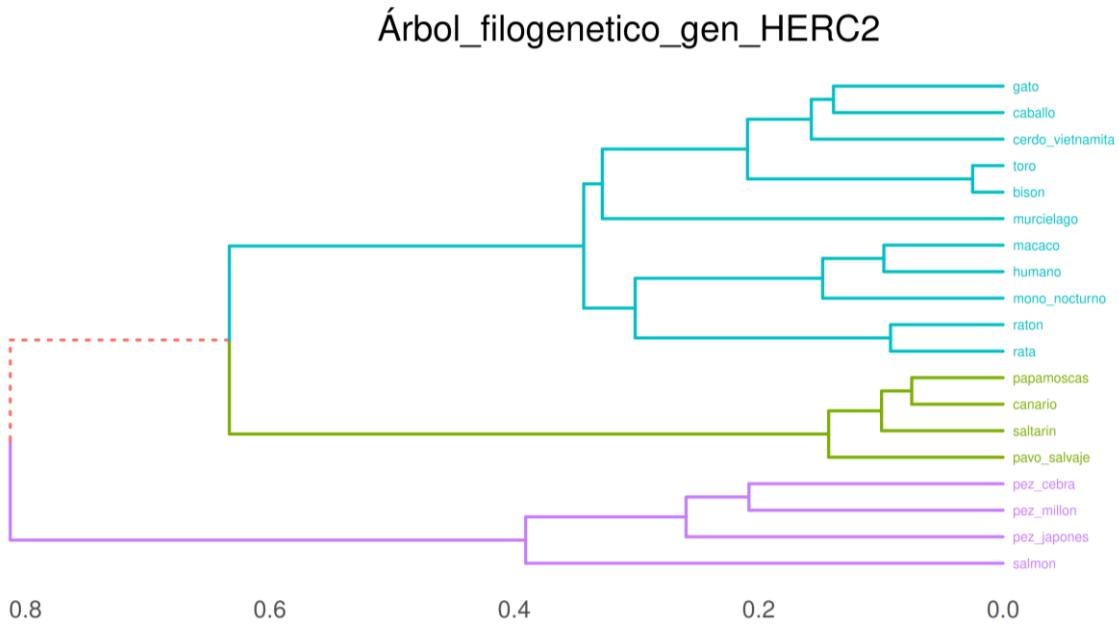
1) Alineamiento de secuencias

- Heatmap



En el heatmap bidimensional se representa en cada rectángulo la distancia relativa entre las especies. Cuanto mayor es dicha distancia, más intenso es el color azul. Nótese por ejemplo la diferencia entre el salmón y el macaco o el humano. Sin embargo, cuando las distancias se acortan el color se torna rojizo, por ejemplo, entre el macaco y el humano.

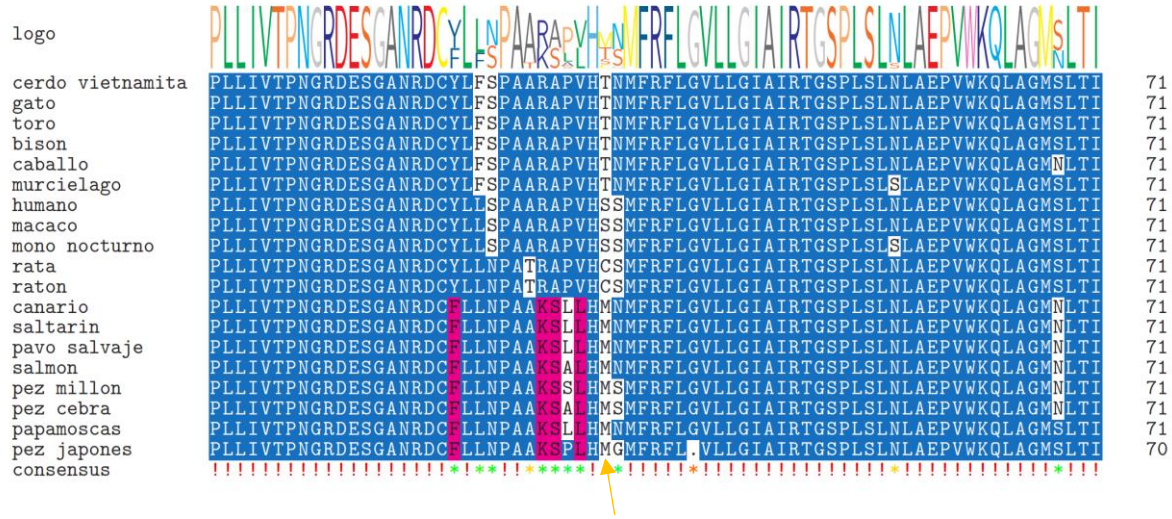
- Árbol filogenético



En el árbol filogenético superior se muestran las especies agrupadas en 3 clados. En cada clado se encuentran las especies agrupadas dependiendo de la clase a la que pertenecen (mamíferos, aves, o peces)

2) Análisis del alineamiento (dominio HECT)

- Alineamiento y secuencia logo

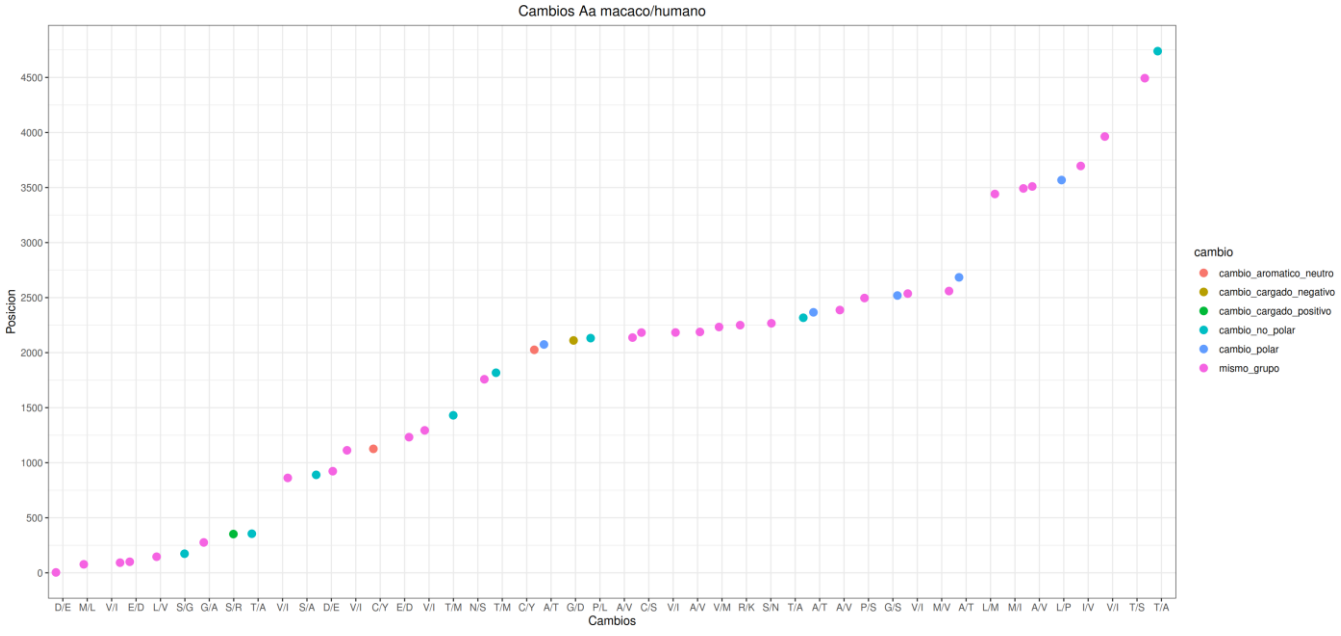


En la región (4540:4610) del dominio HECT existen 4 posiciones (marcadas con rojo) donde se observa un cambio de aminoácido específico entre los mamíferos por una parte y los peces y aves por otra. Así mismo con la flecha naranja se muestra una posición en la que existe un aminoácido específico para los peces y aves (M), otro para los roedores (C), otro para los primates (S) y otro para el resto de los mamíferos (T).

Como comentamos anteriormente, el dominio HECT está implicado en la unión de diversos substratos para su ubiquitinación. Por lo tanto, resultaría interesante realizar experimentos de mutagénesis dirigida para ver si la presencia de un aminoácido concreto (o un grupo de ellos) tiene influencia sobre la función de la proteína HERC2 en una especie determinada. De este modo se facilitaría la comprensión completa del mecanismo molecular que impera en dicha proteína.

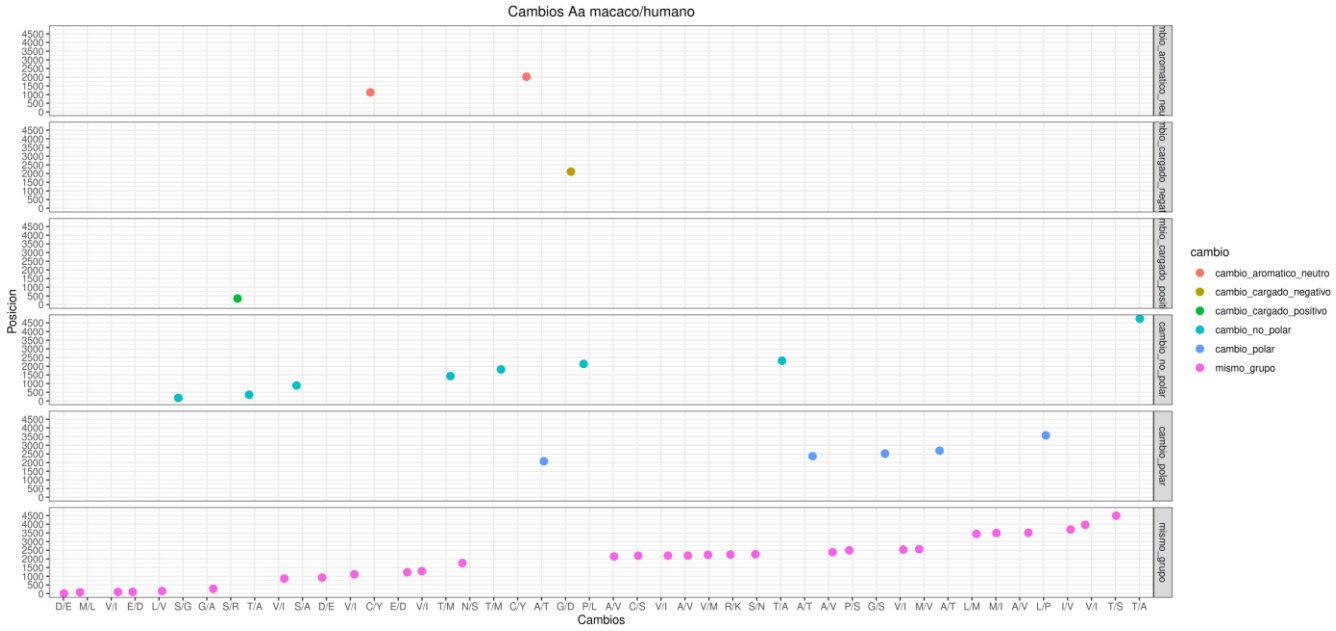
3) Comparativa entre las secuencias de humano y macaco:

- Cambios de aminoácidos (representación con ggplot2)



Entre las secuencias HERC2 de humano y macaco existen en total 46 posiciones donde se producen cambios. Como se muestra en la gráfica superior, la mayor parte de los cambios se producen por aminoácidos del mismo grupo (color rosa).

Finalmente se muestra desglosado con la función facet_grip para visualizarlo mejor



BIBLIOGRAFÍA

1. Tsan-Yuk Lam, T., & Chung-Chau Hon, J.W.T. (2010). Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical Reviews in Clinical Laboratory Sciences 47*(1), 5-49.

2. Sánchez-Tena, S., Cubillos-Rojas, M., Schneider, T., Rosa. J.L. (2016). Functional and pathological relevance of HERC family proteins: a decade later. *Cellular and Molecular Life Sciences 73*(10), 1955-1968

3. Cubillos-Rojas, M., Amair-Pinedo, F., Peiró-Jordán, R., Bartrons, R., Ventura,F., Rosa, J.L. (2014). The E3 Ubiquitin Protein Ligase HERC2 Modulates the Activity of Tumor Protein p53 by Regulating Its Oligomerization. *Journal of Biological Chemistry 289*(21), 14782–14795