



Pontificia Universidad
Católica del Ecuador

FACULTAD DE INGENIERÍA
MAESTRÍA DE BIOLOGÍA COMPUTACIONAL

TRABAJO DE TITULACIÓN

TEMA:

“Estudiar la resistencia al cáncer en algunas especies de aves por medio del análisis de genes codificantes y secuencias conservadas e identificar rutas biológicas asociadas a la baja prevalencia del cáncer”

AUTORA

JOCELYNE CRISTINA PEPINÓS VALLE

DIRECTOR

Ph.D. Daniel Chávez

2023-2024

ÍNDICE DE CONTENIDOS

INTRODUCCIÓN	3
METODOLOGÍA	5
RESULTADOS	8
DISCUSIÓN	19
CONCLUSIÓN	23
REFERENCIAS	23
ANEXOS.....	27

INTRODUCCIÓN

El cáncer constituye una enfermedad que afecta a todo tipo de especies multicelulares desde los animales más pequeños y grandes hasta el ser humano (Kapsetaki et al., 2023). La alta prevalencia de cáncer y la necesidad de mejorar el diagnóstico y las terapias de personas con este padecimiento requieren de mayor investigación científica (Hanna et al., 2021). El cáncer es la segunda causa de mortalidad a nivel mundial, en donde un 70% de los decesos se registran en países en vías de desarrollo o escasos ingresos (Ruales & Checa, 2018). Es por ello, que requiere de especial interés para entenderlo mejor y en un futuro mejorar los análisis y tratamientos.

El principal problema del estudio del cáncer es su gran heterogeneidad y complejidad (Nature Cancer, 2020), lo que hace necesario incrementar su estudio y entendimiento. Una alternativa para ampliar la información sobre el cáncer es la identificación de patrones comunes entre especies como genes y rutas biológicas para poder comprender su desarrollo y su baja prevalencia en ciertos organismos. Se ha observado que las tasas de aparición del cáncer varían dependiendo de las especies estudiadas. De esta manera, Compton et al. (2023) y Vincze et al. (2022) han estudiado la prevalencia de cáncer entre vertebrados y más específicamente entre mamíferos, respectivamente. Compton et al. (2023) generó una base de datos con los registros de necropsias de centros de albergues de animales y al considerar los índices de neoplasia (tumores benignos) y malignidad (tumores malignos) de los vertebrados, se identificó que las aves son un grupo que presenta un bajo índice de neoplasia y malignidad.

Se estudian la prevalencia de neoplasia y malignidad como indicadores del potencial o completo desarrollo del cáncer. Con respecto a la neoplasia, se la define como el crecimiento anormal de células no cancerosas, mientras que la malignidad se denomina como la proliferación excesiva de células cancerígenas (Coleman, 2018; Compton et al., 2023). Todo organismo multicelular se encuentra en riesgo de generar estos identificadores del cáncer (tumores benignos o malignos) debido a que existe la probabilidad de que la división celular normal se descontrole y aumenten las mutaciones somáticas, comúnmente debido a cambios en los genes que regulan el ciclo celular como los genes supresores de tumores y protooncogenes (Coleman, 2018). La función de los genes supresores de tumores consiste en detener la división celular o permitir la muerte celular programada, sin embargo, si presentan mutaciones pueden dar lugar a la proliferación excesiva de células y el cáncer. En comparación, los protooncogenes cumplen funciones de mantener y generar nuevas células previniendo un crecimiento descontrolado que al mutar dan origen a la neoplasia y malignidad (Nenclares & Harrington, 2020). Es por ello por lo que una lista de veinte de estos genes reguladores será considerada en el presente proyecto.

Al analizar las tendencias del cáncer en los distintos organismos, se esperaría que los animales más grandes que viven más años, al poseer más células en su cuerpo presentan mayores eventos de división celular necesarias para crecer y conservar su gran tamaño, lo que incrementaría la probabilidad de generar más errores en la replicación del

ADN que podrían dar origen al cáncer. En contraste, se esperaría que los animales más pequeños que viven menos años al necesitar un menor número de células y replicaciones celulares se encontrarían en menor riesgo de generar mutaciones en la división celular, disminuyendo sus probabilidades de desarrollar cáncer (Seluanov et al., 2018; Yagmur Erten et al., 2020). Sin embargo, se ha evidenciado en elefantes y ballenas un bajo índice de cáncer, lo que se podría atribuir a su capacidad de contrarrestar esta enfermedad (Caulin & Maley, 2011). Es así como surge la paradoja de Peto, la cual plantea que las especies de gran tamaño y que viven más tiempo no presentan más riesgo de cáncer que aquellas pequeñas que viven menos como los ratones (Caulin & Maley, 2011). Sin embargo, se han encontrado ciertas limitaciones de esta paradoja al considerar otros factores como tiempo de gestación (Compton et al., 2023). En el presente proyecto, se considerará la paradoja de Peto para dilucidar los patrones del cáncer al estudiar su incidencia en algunas especies de aves, analizando características como su tamaño y longevidad.

Se utilizó la base de datos de Compton et al. (2023) sobre las tasas de neoplasia y malignidad de los vertebrados, entre ellos las aves. Estos datos se utilizaron para identificar las especies de aves de interés del presente estudio basado en la cantidad total de registros y la baja prevalencia de cáncer. Se seleccionó a las aves debido a su resistencia al desarrollo del cáncer (Møller et al., 2017), presentando índices menores de neoplasia y malignidad que otros animales, entre ellos los mamíferos (Compton et al., 2023). Además, se estima que las aves al ser descendientes de los dinosaurios terópodos (Stiller & Zhang, 2019), pudieron haber mantenido los mecanismos de resistencia al cáncer de estos animales gigantes.

Hace 198 millones de años, los antecesores de las aves eran animales grandes de aproximadamente 163 kg. Se sugiere que las aves atravesaron un proceso de miniaturización continua y adaptaciones para desarrollar la habilidad de volar (Benson et al., 2014). A pesar de su tamaño reducido, se estima que podrían haber conservado la capacidad de supresión de tumores de los dinosaurios, lo que actualmente les confiere menor riesgo de cáncer y mayor longevidad en contraste con los mamíferos voladores (Yagmur Erten et al., 2020). Es así como se escogió las aves como el grupo de interés para este proyecto.

En los últimos años, los tratamientos del cáncer se han centrado en cirugías, quimioterapias y radioterapias. Sin embargo, estos métodos no son específicos para destruir solo células cancerosas sino también las células sanas del cuerpo y generan efectos secundarios en los pacientes. Es por ello, que actualmente se van desarrollando nuevas terapias con células madre e inmunoterapias (Debela et al., 2021). Adicionalmente, en un futuro se espera el desarrollo de nuevas alternativas para tratar estas enfermedades como la edición genética utilizando la herramienta CRISPR (Liu et al., 2023), la cual podría ser más específica, personalizada y eliminar los efectos adversos. De esta manera, el presente proyecto permitiría la identificación de mutaciones únicas que confieran resistencia al cáncer, las cuales podrían ser fundamentales para ser usadas como objetivos (targets) de CRISPR y mejorar los tratamientos actuales del cáncer.

La falta de identificación de patrones de prevalencia del cáncer entre especies de aves motivó al análisis de sus genes codificantes y regiones conservadas descargadas de

la base de datos del NCBI. Se consideran las partes codificantes debido a su rol en la síntesis de proteínas. El estudio principalmente se enfoca en detectar las mutaciones nuevas y únicas en los genes relevantes al cáncer que pueden estar bajo selección positiva en ciertas especies de aves. Para ello se desarrolló completamente un script en la terminal de Linux para realizar automáticamente todo el análisis de distintos protooncogenes y genes supresores de tumores en especies de aves ortólogas. Se consideraron especies ortólogas debido a que provienen de un ancestro común, y a pesar de que divergen entre ellas por medio de la especiación, los genes conservan su función principal (Fernández et al., 2019). Es así, como se podrán comparar estos hallazgos en aves con el ser humano, para potenciales aplicaciones en pacientes con cáncer. El estudio comprendió análisis filogenéticos y la detección de posibles rutas biológicas que impidan el desarrollo del cáncer tales como un mejor sistema inmune, reparación del ADN o apoptosis.

METODOLOGÍA

El estudio de la resistencia al cáncer en algunas especies de aves con el fin de detectar mutaciones únicas en genes codificantes, se lo realizó desarrollando el script llamado “DESCIFRANDO_CANCER_AVES” en la terminal de Linux. Por medio de este pipeline fue posible evaluar de manera automática cada gen seleccionado y determinar aquellos que podrían haber sido sujetos a selección natural. El script completo se lo puede visualizar en el Anexo 1 y en el siguiente enlace:

<https://colab.research.google.com/drive/1PptrUdj4vk9Z-2XhW05GQNYE9Fb0b6DL?usp=sharing>

El script consistió en la descarga de genes y especies de aves seleccionadas del NCBI utilizando las herramientas de la línea de comandos de NCBI Datasets. Adicionalmente, se usaron tres programas bioinformáticos añadidos en el script:

- MUSCLE (Multiple Sequence Comparison by Log- Expectation o Comparación de secuencias múltiples por expectativa logarítmica) para la alineación de las secuencias descargadas.
- RAxML (Randomized Axelerated Maximum Likelihood o Probabilidad Máxima Acelerada Aleatoria) para generar árboles filogenéticos de los datos con máxima verosimilitud.
- HyPhy (Hypothesis Testing using Phylogenies o Prueba de hipótesis utilizando filogenias) para evaluar las secuencias e identificar si la selección natural actúa sobre algunos genes codificantes.

Selección de especies y genes de interés

Primeramente, se seleccionó el grupo de estudio enfocado en las aves debido a su baja prevalencia de cáncer. Se utilizó el artículo sobre la Prevalencia del cáncer en los vertebrados (Compton et al., 2023) para seleccionar el orden de aves con mayor registros

totales y que presentaban tasas de neoplasia y malignidad. De esta manera, el orden Psittaciformes fue seleccionado para el análisis.

En cuanto a la selección de los genes de interés, se consideraron un total de veinte genes entre ellos protooncogenes y genes supresores de tumores registrados en el NCBI y algunos mencionados en artículos científicos de autores como Ulhas Nair et al. (2022), Tejada-Martinez et al. (2021) y Seim et al. (2013).

Para la ejecución de todo este proyecto se utilizó Linux en la máquina virtual Oracle VM VirtualBox, donde se generó el script “DESCIFRANDO_CANCER_AVES” para realizar el estudio de cada gen automáticamente.

Descargar el conjunto de datos de las especies de aves y genes de interés del NCBI

Para realizar la descarga de las especies que presentan los genes de interés, primero se instaló el entorno de anaconda en el terminal de Linux utilizando el siguiente enlace <https://docs.anaconda.com/free/anaconda/install/linux/> de la página de Anaconda. Luego se realizó la instalación del paquete de ncbi_datasets (<https://www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/>) de la página del NCBI.

En la terminal, se usó la línea de comando indicada a continuación para descargar el conjunto de datos (datasets) basado en el símbolo del gen obtenido del NCBI, se incluyeron solo las regiones codificantes y el grupo ortólogo de aves a estudiar “Psittaciformes”.

```
datasets download gene symbol src --include cds --ortholog psittaciformes --filename practiceSRC.zip
```

- Las partes en negrita pueden ser modificadas según el análisis que se solicite

Se esperaba obtener las secuencias de los genes ortólogos debido a que se busca identificar los genes homólogos que son conservados funcionalmente a lo largo de la evolución, en donde los genes presentan un ortólogo en otras especies (Zahn-Zabal et al., 2020).

Alineamiento de secuencias con MUSCLE

MUSCLE es un programa bioinformático que permite realizar alineamientos múltiples de las secuencias de proteínas o nucleótidos. Se caracteriza por su alta precisión y velocidad al alinear una amplia cantidad de secuencias (Edgar, 2004).

En el presente proyecto, se utilizó MUSCLE para realizar la alineación de las secuencias de las especies de aves pertenecientes al orden Psittaciformes previamente descargadas del NCBI. MUSCLE permitió la comparación de las cadenas de nucleótidos de cada especie para posteriormente identificar relaciones evolutivas entre los genes seleccionados. De esta manera, se esperaba visualizar mutaciones únicas presentes en estas especies de aves que les pueda conferir mecanismos de resistencia al cáncer.

Se utilizó la siguiente línea de comando para correr MUSCLE en el terminal:

```
muscle -in SRC-cds-practice/ncbi_dataset/data/cds.fna -out SRCmuscleAlignPrac.afa
```

- -in: Archivo/ruta de entrada en formato FASTA
- -out: Alineación de salida en formato FASTA

Generación de árboles filogenéticos con RAxML

RAxML es un programa bioinformático que permite construir y analizar las filogenias de una amplia cantidad de datos de máxima verosimilitud. Entre la principal ventaja de este algoritmo está su velocidad buscando árboles filogenéticos con puntuaciones de probabilidad altas (Stamatakis, 2014).

Se utilizó RAxMLHPC (versión estándar) para realizar el análisis filogenético de las secuencias de aves previamente alineadas con MUSCLE y poder visualizar en el árbol las relaciones evolutivas y ascendencia común entre las especies estudiadas.

La siguiente línea de comando se usó para generar el análisis en RAxML:

```
raxmlHPC -s /home/manager/Scripts/SRCmuscleAlignMod2.afa -n raxml_trees -m  
GTRCAT -f a -x 123 -p 256 -N autoMRE
```

- -s: Especificar el nombre del archivo de datos de alineación
- -n: Especificar el nombre del archivo de salida
- -m: Modelo de nucleótido GTRCAT (GTR + Optimización de tasas de sustitución + Optimización de tasas evolutivas específicas del sitio)
- -f: seleccionar algoritmo
- -x: Especificar un número entero (semilla aleatoria) y activa el arranque rápido
- -p: Especificar una semilla de número aleatorio para las inferencias de parsimonia para reproducir sus resultados y depurar el programa.
- -N: Especificar el número de ejecuciones alternativas en distintos árboles iniciales.

Evaluando la selección natural con HyPhy

HyPhy (prueba de hipótesis que utiliza filogenias) constituye un paquete de software basado en la colaboración abierta. Permite analizar alineamientos múltiples de secuencias usando distintos modelos evolutivos para datos de proteínas, nucleótidos, entre otros. Se obtienen estimaciones de parámetros y pruebas de hipótesis para determinar aspectos como selección natural, tasas evolutivas, recombinación y coevolución (Kosakovsky Pond et al., 2020). HyPhy fue instalado con Conda en la terminal siguiendo las instrucciones de la página oficial (<https://hyphy.org/installation/>).

Hyphy fue utilizado para identificar las especies de aves que contienen genes únicos asociados a la resistencia del cáncer que se encuentran bajo selección natural considerando los p-valores obtenidos. Los p-valores significativos serán las especies con selección natural positiva. Se utilizó el método aBSREL (una probabilidad adaptativa de efectos aleatorios de sitio de sucursal) inicialmente para identificar las ramas y nodos bajo selección positiva (Smith et al., 2015), luego se usó el método MEME (Modelo de evolución de efectos mixtos) para determinar los sitios/codones específicos que se encontraban bajo selección (Murrell et al., 2012).

El siguiente comando se usó para generar los análisis de HyPhy:

```
hyphy LIBPATH=/home/manager/anaconda3/lib/hyphy absrel --alignment  
/home/manager/Scripts/SRCmuscleAlignMod2.afa --tree  
/home/manager/Scripts/RAxML_bestTree.raxml_trees --branches All --output  
/home/manager/Scripts/hyphySRC.json
```

- LIBPATH: ruta de la ubicación de hyphy
- aBSREL o MEME son los métodos utilizados para inferir la presión de selección
- --alignment: alineamiento generado con muscle
- --tree: árbol filogenético generado con raxml
- --branches: ramas para analizar
- --output: especificar la ruta y el nombre del archivo de salida

Análisis de enriquecimiento de los genes seleccionados

Se utilizó el servidor web g:Profiler (<https://biit.cs.ut.ee/gprofiler/gost>) para realizar los estudios del enriquecimiento funcional del conjunto de genes seleccionados que comprendían protooncogenes y genes supresores de tumores. G:Profiler se distingue por proporcionar información actualizada y de alta calidad de una variedad de genes/proteínas y organismos. Raudvere et al. (2019) señala que g:Profiler posee diversas herramientas para el análisis de genes o proteínas tales como g:GOST (Perfil funcional), g:Convert (Conversión de ID del gen), g:Orth (Búsqueda de ortología) y g:SNPense (ID de SNP al nombre del gen). En este proyecto, se utilizó la herramienta g:GOST para poder determinar las posibles rutas biológicas asociadas con la baja prevalencia al cáncer de la lista de genes proporcionada en el sistema.

RESULTADOS

Descripción general

El pipeline “DESCIFRANDO_CANCER_AVES” fue utilizado para analizar la lista de 20 genes codificantes que incluían 10 proto oncogenes y 10 genes supresores de tumores. Al descargar las secuencias se obtuvo un archivo en formato FASTA para cada gen con las respectivas secuencias de aves del orden Psittaciformes. En la tabla 1, se puede observar el resumen de las secuencias descargadas de cada gen. Posteriormente, se generó otro archivo de formato .afa para cada gen con los alineamientos de las secuencias después de utilizar MUSCLE. Se realizaron modificaciones adicionales en cada archivo y se obtuvo un nuevo documento .afa libre de codones stop y solo con el identificador único de la secuencia y las secuencias de ADN. A continuación, se usó RAxML y se generó un archivo para cada gen con el árbol filogenético de las secuencias. Finalmente, se corrió el programa HyPhy con los métodos aBSREL (Smith et al., 2015) y MEME (Murrell et al., 2012) y se obtuvieron dos archivos de formato .json con las ramas y codones bajo selección positiva, respectivamente. En la tabla 2, se encuentra el resumen de los resultados de HyPhy de cada gen. Los archivos generados se visualizaron en HyPhy Vision (<http://vision.hyphy.org/>). Observar en la figura 1

el resumen de la metodología general del script. Ver Anexo 2, las capturas del proceso completo del script corrido en la terminal de Linux para un gen.

Identificación de genes evolucionados bajo selección positiva

Después de analizar alrededor de 35-40 genes protooncogenes y genes supresores de tumores para las especies de aves, se encontraron que 20 genes están presentes en algunas de las aves del Orden Psittaciformes. Los diez primeros genes representan proto oncogenes, estos son: SRC, BRD4, HOXC11, MDM4, KRAS, MYC, JAK2, Akt1, RAF1 y MYB. Los diez siguientes son genes supresores de tumores, los cuales son: SMAD4, RAD17, FAS, ZMYND8, PTEN, ADAMTS8, ANXA1, DAB2, EPHA2 y FBXO31 (Tabla 1).

Al utilizar el método aBSREL con HyPhy, se evidenció que ocho de los 20 genes presentes en el Orden Psittaciformes, estaban bajo selección natural en ciertas proporciones de ramas. Se encontró dos proto oncogenes con ramas bajo selección, SRC y MYB, al tener valores menores a $p=0.0500$. Se identificaron seis genes supresores de tumores cuyas ramas están bajo selección, RAD17, FAS, ZMYND8, PTEN, EPHA2 y FBXO31, los cuales fueron estadísticamente significativos (Tabla 2).

Adicionalmente, se usó el método MEME con HyPhy para identificar en una proporción de ramas los sitios/codones bajo selección positiva (Murrell et al., 2012). Se encontró que 11 de los 20 genes presentaban sitios que pudieron haber evolucionado bajo selección. Cuatro de ellos son proto oncogenes, SRC, MYB, KRAS y JAK2, mientras que los siete restantes son genes supresores de tumores, RAD17, FAS, ZMYND8, EPHA2, FBXO31, ADAMTS8 y DAB2 (Tabla 2).

RESUMEN DE TODOS LOS GENES DESCARGADOS DEL NCBI

N°	GEN ID	NOMBRE GEN	TIPO DE GEN	N° SECUENCIAS CDS DESCARGADAS	RAMAS BAJO SELECCIÓN (ABSREL)	CODONES BAJO SELECCIÓN (MEME)	DESCRIPCIÓN DEL NCBI https://www.ncbi.nlm.nih.gov/	REFERENCIA DEL GEN
1	SRC	SRC proto-oncogene, non-receptor tyrosine kinase	PROTOONCOGEN	14	SI	SI	Rol regulando el desarrollo embrionario y crecimiento celular	NCBI
2	BRD4	Bromodomain containing 4	PROTOONCOGEN	12	NO	NO	Codifica proteínas homólogas asociadas a la mitosis, y a la serina/treonina quinasa	Nair et al. (2022)
3	HOXC11	Homeobox C11	PROTOONCOGEN	4	NO	NO	Codifica proteínas asociadas a la morfogénesis	Nair et al. (2022)
4	MDM4	Regulador MDM4 de p53	PROTOONCOGEN	8	NO	NO	Codifica proteína nuclear que se une a un dominio de p53	Nair et al. (2022)
5	KRAS	KRAS proto-oncogene, GTPase	PROTOONCOGEN	20	NO	SI	Codificación de una proteína de la superfamilia GTPasa.	NCBI
6	MYC	MYC proto-oncogene, bHLH transcription factor	PROTOONCOGEN	5	NO	NO	Apoptosis, rol en el ciclo celular y transformación celular.	NCBI
7	JAK2	Janus kinase 2	PROTOONCOGEN	18	NO	SI	Codifica proteína con rol principal en la señalización de citocinas y factores de crecimiento	NCBI
8	Akt1	AKT1 – AKT serine/threonine kinase 1	PROTOONCOGEN	9	NO	NO	Codifica proteína (Akt) que regula factores de crecimiento	NCBI
9	RAF1	RAF1 – Raf-1 proto-oncogene, serine/threonine kinase	PROTOONCOGEN	22	NO	NO	Codifica la proteína MAP3K, está aguas abajo de la familia Ras de GTPasas	NCBI
10	MYB	MYB proto-oncogene, transcription factor	PROTOONCOGEN	17	SI	SI	Codifica la proteína que se une a ADN HTH (regula la transcripción)	Nair et al. (2022)
11	SMAD4	SMAD family member 4	GEN SUPRESOR DE TUMORES	11	NO	NO	Codifica una proteína Smad de transducción de señales.	Nair et al. (2022)
12	RAD17	RAD17 checkpoint clamp loader component	GEN SUPRESOR DE TUMORES	14	SI	SI	Codifica una proteína que controla el ciclo celular	Nair et al. (2022)
13	FAS	Fas cell surface death receptor	GEN SUPRESOR DE TUMORES	8	SI	SI	Codifica proteínas que regulan la muerte celular programada	Nair et al. (2022)

14	ZMYND8	Zinc finger MYND-type containing 8	GEN SUPRESOR DE TUMORES	35	SI	SI	Una proteína codificada que regula la transcripción	NCBI
15	PTEN	Phosphatase and tensin homolog	GEN SUPRESOR DE TUMORES	8	SI	NO	Supresor de tumores de la ruta AKT/PKB	NCBI
16	ADAMTS8	ADAM metalloproteinase with thrombospondin type 1 motif 8	GEN SUPRESOR DE TUMORES	4	NO	SI	Codifica proteínas ADAMTS, es un posible supresor de tumores	Tejada-Martinez et al. (2021)
17	ANXA1	Annexin A1	GEN SUPRESOR DE TUMORES	5	NO	NO	Codifica proteínas que se encargan de la inhibición de fosfolipasa A2 y tiene capacidad antiinflamatoria	Tejada-Martinez et al. (2021)
18	DAB2	DAB adaptor protein 2	GEN SUPRESOR DE TUMORES	11	NO	SI	La fosfoproteína codificada que responde a mitógenos. Rol como supresor de tumores	Tejada-Martinez et al. (2021)
19	EPHA2	EPH receptor A2	GEN SUPRESOR DE TUMORES	4	SI	SI	Codificación de proteínas tirosina quinasa. Rol en el desarrollo del sistema nervioso	Tejada-Martinez et al. (2021)
20	FBXO31	F-box protein 31	GEN SUPRESOR DE TUMORES	12	SI	SI	codificación de proteínas de la familia F-box. Rol en la regulación del ciclo celular	Seim et al. (2013)

Tabla 1. Listado de genes descargados del NCBI. Se consideraron veinte genes con un rol en el desarrollo del cáncer que se encuentran presentes en Aves Psittaciformes. Los diez primeros en color amarillo representan protooncogenes y los diez siguientes en naranja son genes supresores de tumores. En la tabla se presenta el ID de cada gen y su respectiva descripción. Se muestra el número de secuencias codificantes que se descargaron en la terminal de Linux de cada gen usando el comando de ncbi_datasets. Se indica Si/No si se obtuvo evidencia de selección positiva con cada uno de los dos métodos de HyPhy, aBSREL (p-value=0.05) y MEME (p-value<=0.1). La referencia del gen representa los artículos y plataformas que previamente mencionaron o estudiaron estos genes relacionados al cáncer. En negrita se encuentran los genes que presentaron selección positiva en ambos métodos (MEME y aBSREL).

RESUMEN DE LOS GENES BAJO SELECCIÓN					
N°	GEN ID	N° RAMAS BAJO SELECCIÓN POSITIVA (aBSREL)	RESULTADOS aBSREL: RAMAS/NODOS Y P- VALUES (p=0.05)	N° SITIOS/CODONES BAJO SELECCIÓN POSITIVA (MEME)	RESULTADOS MEME: CODONES Y P-VALUES (p<=0.1**)
1	SRC	1	<ul style="list-style-type: none"> • XM_010017387_1, p-value = 0.00000 	1	<ul style="list-style-type: none"> • Codon 51, p = 0.0882
2	MYB	4	<ul style="list-style-type: none"> • Node6, p-value = 0.00042 • Node23, p-value = 0.00041 • Node7, p-value = 0.00550 • Node18, p-value = 0.00625 	3	<ul style="list-style-type: none"> • Codon 7, p = 0.0512 • Codon 10, p = 0.0139 • Codon 12, p = 0.0099
3	KRAS	0		4	<ul style="list-style-type: none"> • Codon 57, p = 0.0966 • Codon165, p = 0.0776 • Codon 237, p = 0.0735 • Codon 269, p = 0.0883
4	JAK2	0		1	<ul style="list-style-type: none"> • Codon 199, p = 0.0373
5	RAD17	2	<ul style="list-style-type: none"> • XM_031044362_2, p-value = 0.00000 • XM_010016178_1, p-value = 0.00000 	8	<ul style="list-style-type: none"> • Codon 7, p = 0.0989 • Codon 12, p = 0.0253 • Codon 13, p = 0.0299 • Codon 22, p = 0.0986 • Codon 106, p = 0.0762 • Codon 155, p = 0.0742 • Codon 603, p = 0.0249 • Codon 677, p = 0.0328
6	FAS	3	<ul style="list-style-type: none"> • XM_010009552_1, p-value = 0.00303 • Node3, p-value = 0.01505 • XM_057401095_1, p-value = 0.01830 	16	<ul style="list-style-type: none"> • Codon 95, p = 0.0651 • Codon 102, p = 0.0596 • Codon 103, p = 0.0683 • Codon 188, p = 0.0478 • Codon 445, p = 0.0942 • Codon 480, p = 0.0634 • Codon 866, p = 0.0018 • Codon 1101, p = 0.0824 • Codon 1281, p = 0.0754 • Codon 1355, p = 0.0273 • Codon 1372, p = 0.0209 • Codon 1833, p = 0.0626 • Codon 1961, p = 0.0152 • Codon 2176, p = 0.0102 • Codon 2246, p = 0.0790

					<ul style="list-style-type: none"> Codon 2406, p = 0.0008
7	ZMYND8	8	<ul style="list-style-type: none"> Node19, p-value = 0.00000 Node64, p-value = 0.00000 XM_057432157_1, p-value = 0.00001 XM_057432150_1, p-value = 0.00004 XM_030503264_1, p-value = 0.00004 XM_013128472_3, p-value = 0.00004 XM_030503269_1, p-value = 0.00094 XM_034066649_1, p-value = 0.01825 	11	<ul style="list-style-type: none"> Codon 22, p = 0.0205 Codon 1179, p = 0.0060 Codon 1207, p = 0.0005 Codon 1209, p = 0.0002 Codon 1210, p = 0.0215 Codon 1214, p = 0.0825 Codon 1215, p = 0.0400 Codon 1223, p = 0.0370 Codon 1233, p = 0.0084 Codon 1238, p = 0.0286 Codon 1243, p = 0.0494
8	PTEN	1	<ul style="list-style-type: none"> XM_030487348_2, p-value = 0.00003 	0	
9	EPHA2	4	<ul style="list-style-type: none"> XM_034068124_1, p-value = 0.00000 Node2, p-value = 0.00004 XM_030507884_1, p-value = 0.00003 XM_057401309_1, p-value = 0.00423 	3	<ul style="list-style-type: none"> Codon 38, p = 0.0973 Codon 572, p = 0.0757 Codon 597, p = 0.0063 <i>Nestor notabilis</i> XM_010012765_1
10	FBXO31	1	<ul style="list-style-type: none"> Node2, p-value = 0.00003 	11	<ul style="list-style-type: none"> Codon 35, p = 0.0702 Codon 141, p = 0.037 Codon 269, p = 0.0884 Codon 309, p = 0.0084 Codon 322, p = 0.0462 Codon 450, p = 0.0328 Codon 521, p = 0.0934 Codon 530, p = 0.0335 Codon 531, p = 0.0468 Codon 613, p = 0.0758 Codon 640, p = 0.0263
11	ADAMTS8	0		1	<ul style="list-style-type: none"> Codon 366, p = 0.0360
12	DAB2	0		3	<ul style="list-style-type: none"> Codon 208, p = 0.0277 Codon 209, p = 0.0434 Codon 210, p = 0.0000

Tabla 2. Descripción de los resultados de HyPhy de los posibles genes bajo selección positiva. Se presentan las ramas/nodos significativos de cada gen obtenidas con aBSREL que podrían haber evolucionado bajo selección positiva (Holm-Bonferroni corregido p-value=0.05). Los resultados de MEME señalan los sitios/codones significativos que dan indicio de selección positiva en una proporción de las ramas (p-value<=0.1). Subrayado de amarillo se encuentran las ramas y codones de interés que serán explicadas posteriormente.

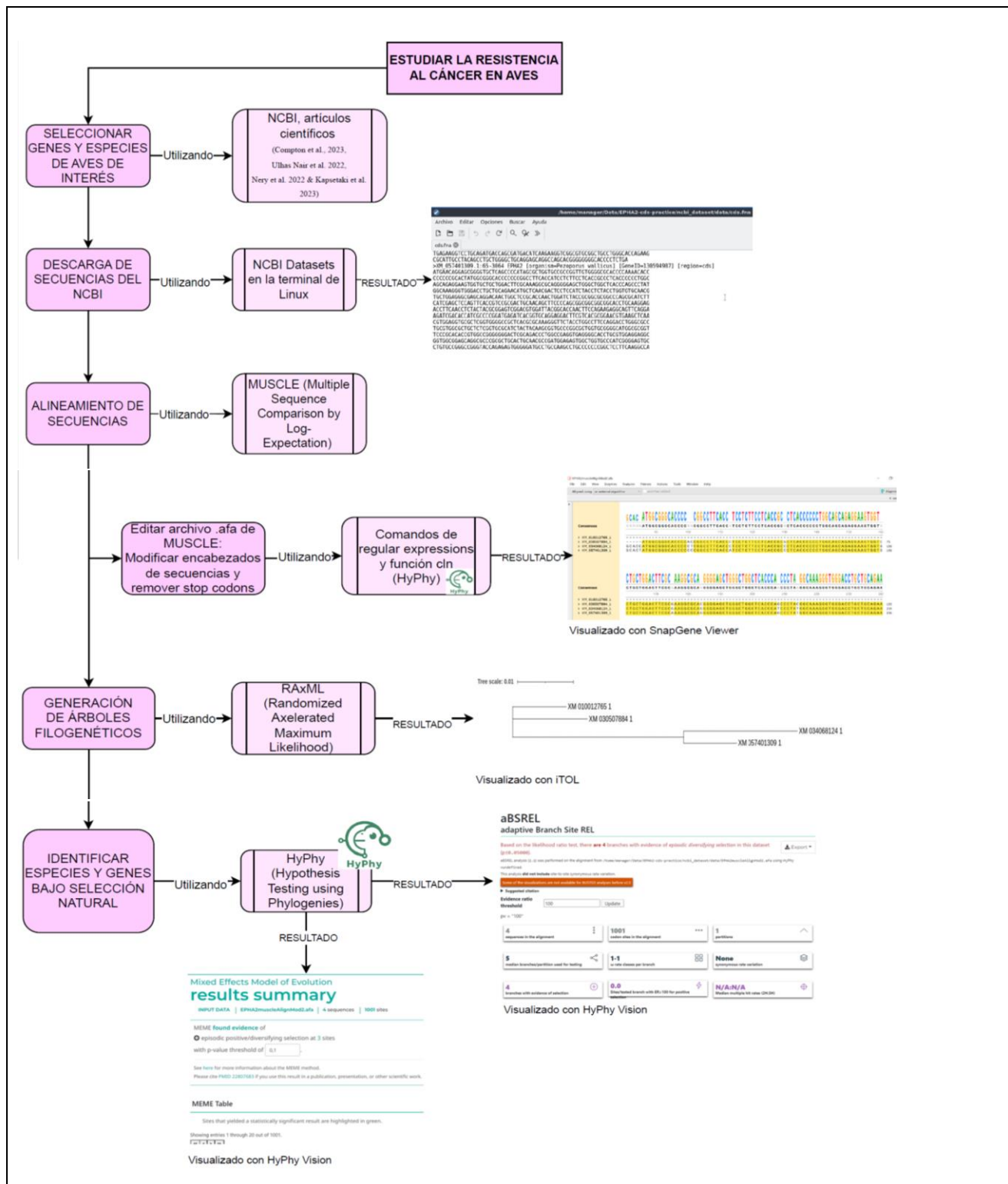


Figura 1. Resumen de la metodología del script y resultados generales. Los genes asociados al cáncer y especies de aves se seleccionaron del NCBI y de artículos científicos. Se usó la herramienta de ncbi_datasets para descargar las secuencias de cada gen en un archivo FASTA individual; MUSCLE permitió alinear las secuencias descargadas. Previamente a utilizar RAxML, se editó el archivo de MUSCLE modificando el encabezado de cada secuencia con comandos de expresiones regulares y se eliminaron los stop codones con la función cin de HyPhy. El archivo generado se visualizó con SnapGene Viewer. RAxML permitió generar filogenias que se visualizaron con iTOL. Al final, se usó HyPhy para obtener los resultados de las ramas (con aBSREL) y de los codones (con MEME) significativos que sugieran selección positiva. Estos archivos se observaron en HyPhy Vision.

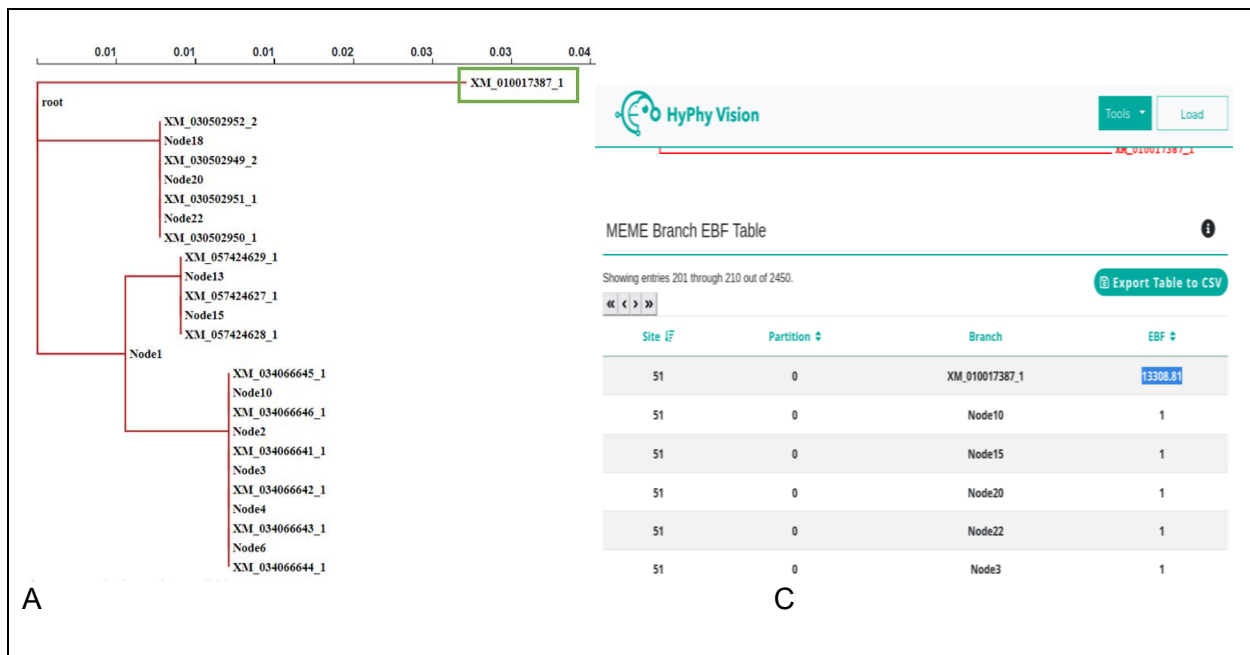
Determinación de genes y sitios de interés de aves Psittaciformes

Los archivos obtenidos con HyPhy fueron visualizados en HyPhy Vision. Al analizar los resultados de MEME, se consideraron los valores EBF (empirical Bayes factor) para explorar los datos y determinar una ubicación potencial que sugiera la evidencia de selección en los sitios/codones en ramas del árbol filogenético obtenido con RAxML para cada gen. Murrell et al., 2012 señaló las limitaciones de MEME al identificar precisamente los sitios y las ramas bajo selección; es por ello que se consideró los resultados de MEME en comparación con los resultados de aBSREL para distinguir las posibles ramas de interés.

Del listado de 20 genes, se encontraron dos candidatos después de comparar los resultados de MEME y aBSREL. Estos genes son el protooncogén SRC (non-receptor tyrosine kinase) y el gen supresor de tumores RAD17 (RAD17 checkpoint clamp loader component).

Con respecto al gen SRC, se identificó en aBSREL la rama con el identificador de secuencia XM_010017387_1 con un p valor significativo (se estimó $p=0.0500$). De la misma manera, en MEME se encontró el codón 51 bajo selección positiva al ser significativo ($p<=0.1$). Este sitio podría ocurrir en la rama XM_010017387_1 al considerar el valor EBF más alto. Observar en la figura 2 el árbol filogenético de este gen y los valores EBF del sitio y rama seleccionadas.

Al considerar el gen RAD17, se evidenció la rama con identificador XM_010016178_1 bajo selección positiva en aBSREL ($p=0.0500$). Igualmente, en MEME se encontró que el codón 12 es estadísticamente significativo con un $p = 0.0253$ y su localización podría presentarse en la rama XM_010016178_1 al poseer el mayor valor EBF de todos los sitios identificados por HyPhy. En la figura 3, se puede visualizar los resultados de HyPhy de este gen y de su respectivo árbol filogenético.



```

### For partition 1 these sites are significant at p <=0.1
  Codon | Partition | alpha | beta+ | p+ | LRT | Episodic selection detected? | # branches |
  -----|-----|-----|-----|-----|-----|-----|-----|
    51 | 1 | 0.000 | 32.419 | 0.253 | 3.363 | Yes, p = 0.0882 | 1 |
### ** Found _1_ sites under episodic diversifying positive selection at p <= 0.1**

```

B

Figura 2. Resultados de RAXML y HyPhy del gen SRC. (A) Árbol filogenético generado con RAXML y visualizado en iTOL. También se lo observó en HyPhy Vision con el archivo .json de HyPhy. El recuadro verde es la rama que se estima presenta selección positiva. Las medidas en la parte superior permiten distinguir la distancia entre las ramas. (B) Resultados de HyPhy con MEME mostrados en la terminal. Codón significativo ($p <= 0.1$) que podría señalar selección positiva episódica en una rama. (C) Resultados de HyPhy del sitio/codón significativo visualizados en HyPhy Vision. Se presentan todas las ramas y nodos con sus respectivos valores EBF. La rama XM_010017387_1 posee el valor EBF más alto en este sitio (EBF=13308.81).



A

C

```

### For partition 1 these sites are significant at p <=0.1
  Codon | Partition | alpha | beta+ | p+ | LRT | Episodic selection detected? | # branches |
  -----|-----|-----|-----|-----|-----|-----|-----|
    7 | 1 | 0.000 | 33.690 | 0.199 | 3.145 | Yes, p = 0.0989 | 1 |
   12 | 1 | 0.000 | 67.408 | 0.178 | 5.779 | Yes, p = 0.0253 | 1 |
   13 | 1 | 0.000 | 78.175 | 0.191 | 5.452 | Yes, p = 0.0299 | 1 |
   22 | 1 | 0.000 | 16.978 | 0.261 | 3.152 | Yes, p = 0.0986 | 1 |
  106 | 1 | 0.000 | 25.924 | 0.212 | 3.644 | Yes, p = 0.0762 | 1 |
  155 | 1 | 0.000 | 259.705 | 0.144 | 3.693 | Yes, p = 0.0742 | 1 |
  603 | 1 | 0.000 | 1010.641 | 0.124 | 5.809 | Yes, p = 0.0249 | 1 |
  677 | 1 | 0.000 | 610.889 | 0.165 | 5.270 | Yes, p = 0.0328 | 1 |
### ** Found _8_ sites under episodic diversifying positive selection at p <= 0.1**

```

B

Figura 3. Resultados de RAXML y HyPhy del gen RAD17. (A) Árbol filogenético generado con RAXML y visualizado en iTOL o HyPhy Vision (archivo .json). El recuadro verde representa la rama que se estima presenta selección positiva. En la parte superior, se encuentran las dimensiones para distinguir la distancia entre las ramas. (B) Resultados de HyPhy con MEME obtenidos en la terminal. Codones/sitios significativos ($p <= 0.1$) que podrían indicar selección positiva episódica en una rama. (C) Resultado de HyPhy del sitio significativo, visualizado en HyPhy Vision. Se presentan todas las ramas y nodos con sus respectivos valores EBF. La rama XM_010016178_1 posee el valor EBF más alto en este sitio (EBF=3374.3).

Detección de las especies de aves correspondientes a las ramas bajo selección de los dos genes de interés

Una vez se determinaron las posibles ramas de los árboles filogenéticos bajo selección positiva de los genes SRC y RAD17, se analizó sus archivos FASTA descargados en la terminal para identificar el organismo al que corresponde cada identificador de secuencia del árbol. Se encontró que tanto la rama XM_010017387_1 del gen SRC como la

rama XM_010016178_1 del gen RAD17 pertenecen al organismo *Nestor notabilis* (Figura 4). Este organismo representa una especie de ave del orden Psittaciformes caracterizada por su gran tamaño y longevidad (Aitken et al., 2023; Pullar, 1996).



Figura 4. Identificación de las especies de aves correspondientes a las ramas seleccionadas. Archivos .fna de las secuencias codificantes del orden de aves Psittaciformes de los genes más significativos encontrados previamente (SRC y RAD17). (A) Rama XM_010017387_1 del gen SRC del recuadro rojo representa a la especie *Nestor notabilis*. (B) Rama XM_010016178_1 del gen RAD17 del recuadro rojo corresponde al ave *Nestor notabilis*.

Localización de mutaciones únicas en los genes de interés (SRC y RAD17)

Al analizar las secuencias de nucleótidos del gen SRC, se identificó que la secuencia XM_010017387_1 presenta diferencias en el codón 51 en contraste con las otras 13 secuencias de los organismos que también presentan este gen. Como fue previamente mencionado, el codón 51 sugiere evolución bajo selección positiva el cual se asoció con la secuencia XM_010017387_1. Se observó en la figura 5 con SnapGene Viewer, la presencia triple de Adeninas (AAA) en el codón 51 (posición 151-153 bases), en lugar del codón CCT encontrado en las demás secuencias descargadas. Además, se observa una inserción de dos Adeninas (AA) adicionales delante del codón 51 que difiere de las otras secuencias.

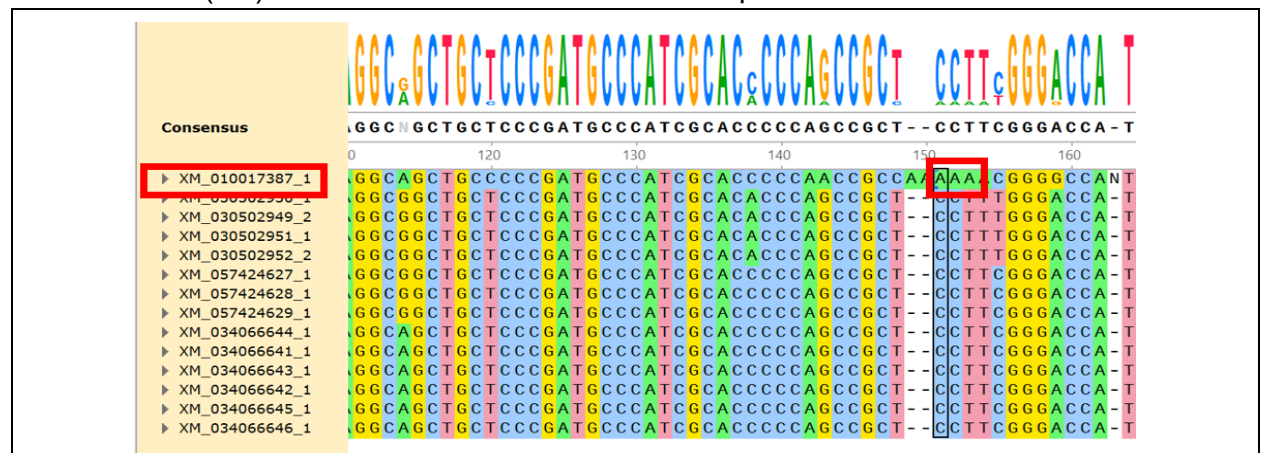


Figura 5. Mutaciones únicas en la secuencia XM_010017387_1 del gen SRC. El recuadro rojo de la izquierda representa la secuencia de interés con posible evidencia de selección positiva identificada con aBSREL e inferida de MEME. El recuadro de la derecha es el codón 51 encontrado con MEME, presenta una mutación única (AAA) característica de esta secuencia.

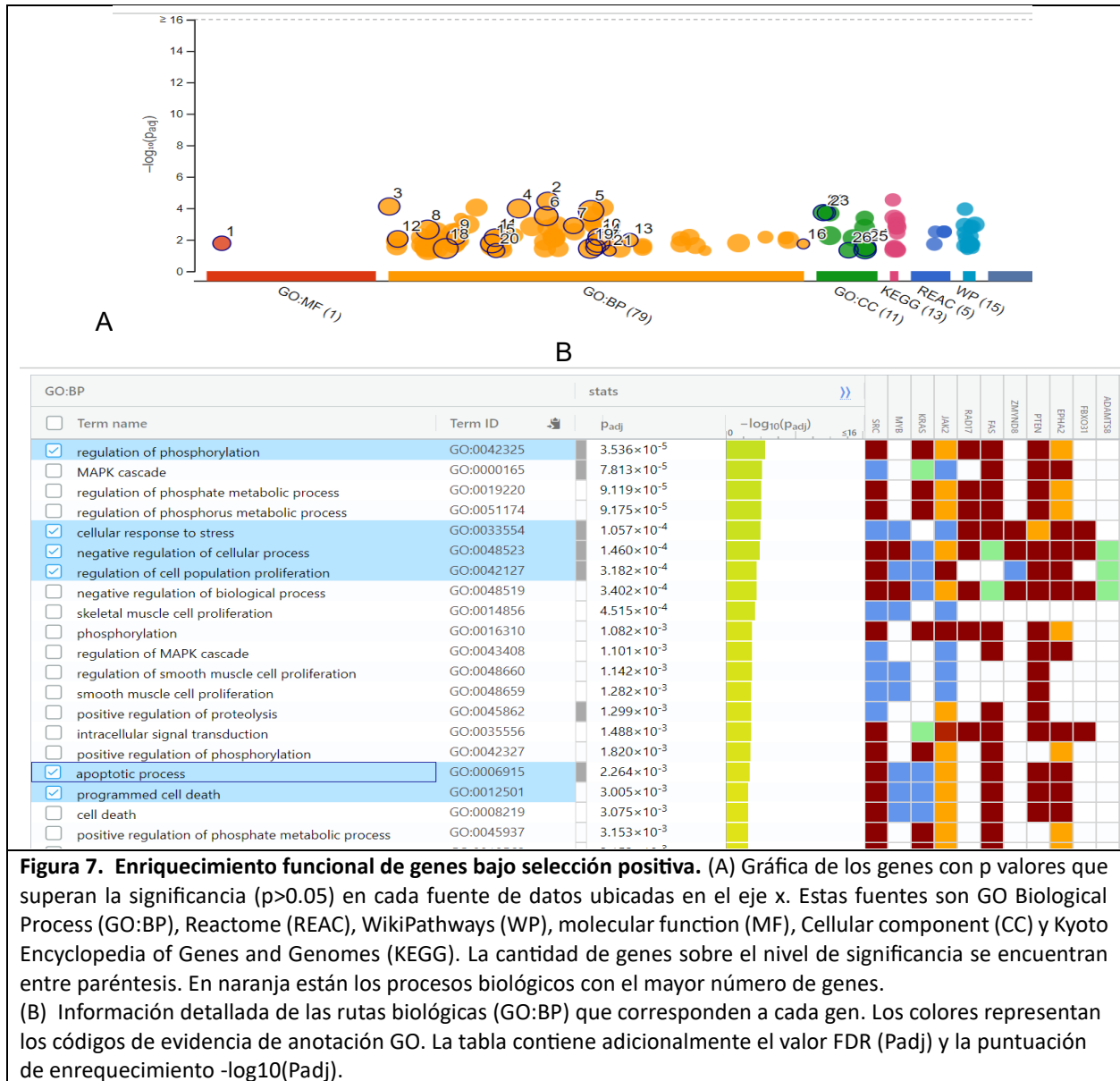


Figura 7. Enriquecimiento funcional de genes bajo selección positiva. (A) Gráfica de los genes con p valores que superan la significancia ($p > 0.05$) en cada fuente de datos ubicadas en el eje x. Estas fuentes son GO Biological Process (GO:BP), Reactome (REAC), WikiPathways (WP), molecular function (MF), Cellular component (CC) y Kyoto Encyclopedia of Genes and Genomes (KEGG). La cantidad de genes sobre el nivel de significancia se encuentran entre paréntesis. En naranja están los procesos biológicos con el mayor número de genes. (B) Información detallada de las rutas biológicas (GO:BP) que corresponden a cada gen. Los colores representan los códigos de evidencia de anotación GO. La tabla contiene adicionalmente el valor FDR (Padj) y la puntuación de enriquecimiento $-\log_{10}(\text{Padj})$.

DISCUSIÓN

Genes evolucionados bajo selección positiva con MEME y aBSREL

Al evaluar distintos genes que tienen un rol significativo en el desarrollo del cáncer, se identificó que la lista de 20 genes de la tabla 1 presentan secuencias disponibles de especies de aves Psittaciformes para descargar del NCBI. Por lo tanto, estos genes se consideraron como candidatos que podrían encontrarse en estas aves. La significancia de cada gen se evaluó posteriormente con los resultados de HyPhy; los genes significativos se resumen en la tabla 2. Algunos de los genes significativos han sido previamente estudiados

en vertebrados por otros autores. Ulhas Nair et al. (2022) realizaron la caracterización de varios genes que presentan una correlación positiva o negativa con la baja prevalencia de cáncer en peces, mamíferos y aves. De este estudio, se seleccionó el protooncogén MYB y los genes supresores de tumores RAD17 y FAS que fueron significativos en HyPhy y podrían haber evolucionado bajo selección positiva. Se ha evidenciado que MYB se encarga de la regulación de células madre y progenitoras de la médula ósea, el colon y regiones del cerebro. Un alta expresión de este oncogén puede dar lugar al cáncer (Mittra, 2018). Mientras que RAD17 y FAS se han encontrado que representan puntos de control para la reparación de la ruptura del ADN deteniendo el ciclo celular o generando apoptosis (Strasser et al., 2009; Wang et al., 2014). Adicionalmente, se escogieron genes supresores de tumores identificados con evidencia de selección positiva en cetáceos, entre ellos están ADAMTS8, DAB2 y EPHA2 (Tejada-Martinez et al., 2021). Estos genes también fueron significativos en HyPhy (ver tabla 2). EPHA2 resultó significativo para ambos métodos de HyPhy, MEME y aBSREL. Este gen es fundamental regulando la aparición de tumores, su invasión y progresión del cáncer por medio de mecanismos de señalización celular para detectar y combatir las células cancerígenas (Wilson et al., 2021). Además, en cuanto al gen FBXO31 seleccionado para este proyecto que presentó significancia, se encontró que los murciélagos los cuales son mamíferos voladores y longevos poseen copias extras de este gen que es importante en el ciclo celular y en la reparación del ADN (Seim et al., 2013).

Se ha encontrado que las aves presentan bajos índices de cáncer (Møller et al., 2017), lo que podría asociarse con su tendencia a vivir por largos periodos de tiempo dependiendo de la especie de ave. Compton et al. (2023) indica que las aves exhibieron un menor porcentaje de los indicadores del cáncer como son la neoplasia y la malignidad. Para comprender la capacidad de resistencia al cáncer en las aves y si presentan genes que estuvieron bajo selección positiva, se consideraron los resultados de HyPhy. Se utilizó las medidas del factor empírico de Bayes (EBF) de cada rama obtenidas con MEME para realizar análisis exploratorios de los datos y determinar la ubicación potencial de un codón con selección. Sin embargo, no se estiman asociaciones fuertes con estos valores, debido a que no se infiere precisamente que la selección se evidencie en un sitio específico de una rama determinada. Es por ello por lo que se utilizó adicionalmente aBSREL para estimar las posibles ramas evolucionadas bajo selección. aBSREL constituye un método adecuado para identificar ramas individuales con selección. Smith et al. (2015) menciona las ventajas de aBSREL referente a su rendimiento computacional y estadístico, además de su mayor velocidad y capacidad de analizar varias secuencias.

Algunas posibles limitaciones de los presentes resultados podrían asociarse al tipo de análisis que tiende a ser más exploratorio debido a los datos utilizados, como los valores EBF y el análisis de las ramas seleccionadas por el programa, en lugar de ser establecidas con anterioridad. Álvarez-Carretero et al. (2023) menciona que en estudios con pruebas de hipótesis como HyPhy se esperaría generar una hipótesis nula previamente al análisis de los datos para evaluarla. De esta manera, al no especificar ramas para analizar, el programa realiza la búsqueda de los linajes que podrían tener sitios bajo selección positiva, en este caso la hipótesis nula se derivaría de los datos. En futuros estudios, se podría plantear la hipótesis nula con anterioridad sobre ramas específicas y se podría considerar el uso adicional de herramientas como CODEML o EasyCodeML de paquetes actuales como PAML (Phylogenetic Analysis by Maximum Likelihood) que también permiten evaluar selección positiva. Estas herramientas permitirían reforzar o contrastar los resultados de HyPhy y además podrían incluir mejoras en cuanto a la facilidad de uso y evaluación de las hipótesis,

así como, el estudio de selección en diversos genes de un grupo de taxones (Álvarez-Carretero et al., 2023; Gao et al., 2019).

Genes de interés de aves Psittaciformes y mutaciones únicas

Se identificaron dos genes de interés relacionados al cáncer, SRC y RAD17, después de comparar los resultados de HyPhy con MEME y aBSREL de la lista de posibles genes bajo selección positiva (tabla 2). En cuanto al gen supresor de tumores RAD17, su función consiste en actuar como sitio de control de la replicación e identificar las rupturas en el ADN (Wang et al., 2014). Al observar los alineamientos de MUSCLE de las secuencias de ADN de aves Psittaciformes para RAD17, se identificó en el codón 12 de la rama XM_010016178_1 perteneciente a la especie *Nestor notabilis* la presencia de una mutación única (figura 3). La secuencia presenta un codón 12 que inicia con Adenina y Timina (ATG) en contraste con las otras secuencias. Esto podría indicar un sitio de interés que sugiera una explicación potencial de la resistencia al cáncer. Sin embargo, existen ciertas limitaciones con esta inferencia debido a que el alineamiento visualizado en SnapGene para este gen (figura 6) presenta varios espacios debido a la falta de una amplia secuenciación de los genomas de aves. Stiller & Zhang (2019) resalta la escasez actual de genomas de este grupo de animales y señala la importancia de continuar con las secuenciaciones pendientes. Se espera que en un futuro los análisis de secuencias de aves sean más robustos y sea posible mejorar las limitaciones encontradas con un mayor número y longitud de genomas.

Adicionalmente, el protooncogén SRC presentó mutaciones únicas en el codón 51 de la rama XM_010017387_1 referente a la especie *Nestor notabilis* (figura 2) donde existe la presencia triple de Adeninas (AAA) y en el codón 50 se encontró una inserción de dos Adeninas (AA) (figura 2). El alineamiento múltiple de secuencias de este gen es continuo y no presenta abundantes espacios debido a que las secuencias tienden a ser del mismo tamaño. Con respecto al rol de este gen en el cáncer, estudios previos señalan su importancia en la regulación de las rutas biológicas y metabólicas de las células cancerígenas una vez que es activado, esto se observa en humanos y algunos animales. Su función principal es la transmisión de señales a proteínas para activar su síntesis y regular vías de glucólisis, absorción de glucosa, fosforilación oxidativa, entre otras; generando el crecimiento celular y nutriendo las actividades de las células anormales (Pelaz & Tabernero, 2022; Raji et al., 2023). Con estos resultados, se cumplió con el objetivo de detectar los cambios genéticos únicos de genes relevantes al cáncer como SRC y RAD17 que estaban bajo selección positiva en ciertas especies de aves Psittaciformes.

Evaluando la paradoja de las aves y sus rutas biológicas

De las ramas de interés bajo selección en los genes SRC y RAD17, se identificó que las secuencias de ADN representan a una especie de ave del orden Psittaciformes (cacatúas y loros) conocida como *Nestor notabilis* (figura 4). Esta ave pertenece a la familia Psittacidae (loros de Nueva Zelanda) y el género *Nestor*, su nombre común es kea. Este tipo de loro se distingue por su gran tamaño dentro de su familia y por su longevidad de 47 a 50 años (Aitken et al., 2023; Brouwer et al., 2000; Young et al., 2012). En comparación con los mamíferos que tienen un tamaño similar, las aves tienden a vivir incluso cuatro veces más. Incluso los murciélagos que sobreviven más que los otros mamíferos, viven menos que las aves de masa corporal semejante (Holmes & Ottinger, 2003). En cuanto a las aves del orden

Psittaciformes, se caracterizan por ser más longevas, esto se asocia con una menor incidencia al cáncer como se evidenció con la presencia de los protooncogenes y genes supresores de tumores bajo selección positiva de la tabla 2. Young et al. (2012) estudiaron loros en cautiverio y sugieren que las especies de menor tamaño tienden a vivir menos que las más grandes debido a la falta de selección por mejores defensas contra el cáncer. En general, *Nestor notabilis* se encuentra entre las aves más longevas del orden Psittaciformes y dentro de su familia.

De las especies de mamíferos con altos índices de mortalidad, se ha identificado que más del 20% al 40% de este grupo de animales mueren por causas relacionadas al cáncer (Vincze et al., 2022). Sin embargo, el orden Artiodactyla de mamíferos que está constituido por especies grandes como las ballenas, es el grupo con mayor resistencia al cáncer. Esta evidencia apoya la paradoja de Peto que sugiere que las especies más grandes a pesar de su tamaño y longevidad no tienen un riesgo más elevado de contraer cáncer, que los animales más pequeños menos longevos (Caulin & Maley, 2011). De la misma manera, la especie de ave encontrada *Nestor notabilis* podría apoyar la paradoja de Peto debido a su longevidad y tamaño, además de la presencia de genes relacionados al cáncer como protooncogenes o genes supresores de tumores que le pueden proporcionar resistencia al cáncer. Se espera que esta especie como otras del orden Psittaciformes presenten mecanismos que controlen la aparición de células cancerígenas. Algunos de los genes bajo selección presentes en estas aves (tabla 2), han sido previamente identificados en otras especies animales como se mencionó anteriormente. Por ejemplo, el gen EPHA2 que fue seleccionado positivamente en cetáceos los cuales tienden a ser longevos (Tejada-Martinez et al., 2021), también se encontró en las aves Psittaciformes. Este gen como los otros con selección les podrían conferir a estas aves la capacidad de evadir errores en el proceso de división celular y eludir el cáncer, permitiéndoles que sean más longevas como se establece en la paradoja de Peto.

La capacidad de las aves de generar defensas contra el cáncer podrían ser producto de conservar los mecanismos de los dinosaurios terópodos de los cuales descienden. Estudios señalan que las aves atravesaron un proceso de miniaturización y desarrollaron la habilidad de volar como adaptaciones para sobrevivir (Benson et al., 2014). Sin embargo, es posible que hayan mantenido las defensas de resistencia al cáncer de sus ancestros grandes (Yagmur Erten et al., 2020). Esto podría explicar que las aves tienden a ser más longevas que los mamíferos voladores de dimensiones similares, como se encontró con el ave *Nestor notabilis* que podría presentar mecanismos de supresión del cáncer y es longeva. Además, se esperaría que los genes relacionados al cáncer que fueron significativos bajo selección positiva (tabla 2) fueran heredados de los dinosaurios y les confiera a las aves estudiadas su resistencia contra el cáncer para vivir durante muchos años.

Adicionalmente, al identificar rutas biológicas de los genes bajo selección natural positiva (tabla 2), se determinaron que algunas podrían ser responsables de la supresión del cáncer en las aves Psittaciformes y de su longevidad. Entre estas rutas están la regulación de la fosforilación, la respuesta celular al estrés, la regulación de la proliferación de la población celular, el proceso apoptótico y la muerte celular programada. En las aves, se ha encontrado que la longevidad podría estar asociada a la habilidad de reparar del ADN, procesar el ARN y el estrés oxidativo, así como el progreso del ciclo celular (Wirthlin et al., 2018), las cuales podrían estar en concordancia con los resultados obtenidos en g:Profiler (figura 7). Adicionalmente, Ulhas Nair et al. (2022) señala que la disminución de la prevalencia del cáncer es mediada por genes que regulan el ciclo celular, reparan el ADN y

generan una reacción inmunológica. En comparación, la longevidad y la supresión del cáncer podrían requerir de algunas rutas biológicas similares. En el presente proyecto, también se encontraron rutas que previenen la proliferación excesiva de las células, lo cual es característico del cáncer, por medio de la destrucción de células anormales y regulación de los procesos celulares. A su vez, estas rutas biológicas podrían tener un papel fundamental en aumentar la esperanza de vida de las aves.

CONCLUSIÓN

Se exploró la resistencia al cáncer en las aves del orden Psittaciformes analizando 20 genes codificantes relacionados al cáncer y presentes en este grupo de aves. Se utilizó el script creado para este estudio llamado “DESCIFRANDO_CANCER_AVES”. Estos genes comprendían genes supresores de tumores y protooncogenes de los cuales once fueron significativos en HyPhy y podrían presentar evidencia de selección positiva. De estos genes, se identificaron dos (RAD17 y SRC) que presentaron mutaciones únicas en los codones de la especie de ave *Nestor notabilis* en contraste con las demás secuencias de aves alineadas con MUSCLE. Esta ave también llamada Kea se la caracteriza por un tamaño grande y la capacidad de vivir hasta 50 años. Se espera que la longevidad de estas aves se deba a sus mecanismos de resistencia al cáncer evidenciados con la presencia de los genes bajo selección y de rutas biológicas que permitan reparar el ADN y controlar o corregir errores en el ciclo celular. Estas defensas contra el cáncer podrían haber sido heredadas de sus ancestros los dinosaurios y se mantuvieron a lo largo de los años a pesar de atravesar un proceso de miniaturización.

En el futuro, será importante recopilar información adicional sobre más genes relacionados al cáncer y se espera que puedan ser considerados más genomas de aves no solo en cautiverio sino también en la naturaleza, debido a que actualmente la variedad de aves es limitada en los zoológicos (Møller et al., 2017). Además, es necesario generar más registros de cáncer en animales salvajes ya que por el momento es un reto documentar la edad real del organismo y realizar necropsias para estudiar el cáncer (Vincze et al., 2022). Con mayores estudios se podrá ampliar la búsqueda de mutaciones únicas en las aves que nos permitan entender mejor el desarrollo del cáncer. Como sugieren Seluanov et al. (2018), Entendiendo más a detalle cómo se produce la resistencia al cáncer en las especies animales, podrían proporcionar indicios importantes para generar terapias que contrarresten el cáncer en seres humanos. Así el presente proyecto, será un pilar para detectar mutaciones no identificadas en los genomas de las aves Psittaciformes que podrían en un futuro aumentar los conocimientos del cáncer e inclusive apoyar en terapias genéticas como objetivos (targets) de CRISPR para tratar esta enfermedad (Liu et al., 2023). Adicionalmente, el script desarrollado permitirá continuar con la investigación de genes y especies animales relevantes de manera rápida y automática.

REFERENCIAS

Aitken, J., Paterson, A., Ross, J., Orr-Walker, T., & Young, L. (2023). A preliminary study of kea (*Nestor notabilis*) habitat use and diet in plantation forests of Nelson, New Zealand. *New Zealand Journal of Zoology*, 1–17.

<https://doi.org/10.1080/03014223.2023.2251904>

- Álvarez-Carretero, S., Kapli, P., & Yang, Z. (2023). Beginner's Guide on the Use of PAML to Detect Positive Selection. *Molecular Biology and Evolution*, *40*(4), 1–18. <https://doi.org/10.1093/molbev/msad041>
- Benson, R. B. J., Campione, N. E., Carrano, M. T., Mannion, P. D., Sullivan, C., Upchurch, P., & Evans, D. C. (2014). Rates of Dinosaur Body Mass Evolution Indicate 170 Million Years of Sustained Ecological Innovation on the Avian Stem Lineage. *PLoS Biology*, *12*(5). <https://doi.org/10.1371/journal.pbio.1001853>
- Caulin, A. F., & Maley, C. C. (2011). Peto's Paradox: evolution's prescription for cancer prevention. *Trends in Ecology & Evolution*, *26*(4), 175–182. <https://doi.org/10.1016/J.TREE.2011.01.002>
- Coleman, W. B. (2018). Neoplasia. *Molecular Pathology: The Molecular Basis of Human Disease*, 71–97. <https://doi.org/10.1016/B978-0-12-802761-5.00004-3>
- Compton, Z., Harris, V., Mellon, W., Rupp, S., Mallo, D., Kapsetaki, S. E., Wilmot, M., Kennington, R., Noble, K., Baciú, C., Ramirez, L., Peraza, A., Martins, B., Sudhakar, S., Aksoy, S., Furukawa, G., Vincze, O., Giraudeau, M., Duke, E. G., ... Boddy, A. M. (2023). Cancer Prevalence Across Vertebrates. *BioRxiv*, 2023.02.15.527881. <https://www.biorxiv.org/content/10.1101/2023.02.15.527881v1%0Ahttps://www.biorxiv.org/content/10.1101/2023.02.15.527881v1.abstract>
- Debela, D. T., Muzazu, S. G. Y., Heraro, K. D., Ndalama, M. T., Mesele, B. W., Haile, D. C., Kitui, S. K., & Manyazewal, T. (2021). New approaches and procedures for cancer treatment: Current perspectives. *SAGE Open Medicine*, *9*. <https://doi.org/10.1177/20503121211034366>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Fernández, R., Gabaldón, T., & Dessimoz, C. (2019). *Orthology: definitions, inference, and impact on species phylogeny inference*. 1–17. <http://arxiv.org/abs/1903.04530>
- Gao, F., Chen, C., Arab, D. A., Du, Z., He, Y., & Ho, S. Y. W. (2019). EasyCodeML: A visual tool for analysis of selection using CodeML. *Ecology and Evolution*, *9*(7), 3891–3898. <https://doi.org/10.1002/ece3.5015>
- Hanna, C. R., Boyd, K. A., & Jones, R. J. (2021). Evaluating cancer research impact: lessons and examples from existing reviews on approaches to research impact assessment. *Health Research Policy and Systems*, *19*(1), 1–45. <https://doi.org/10.1186/s12961-020-00658-x>
- Kapsetaki, S. E., Compton, Z., Dolan, J., Shawn, M., Duke, E. G., Harrison, T. M., Aksoy, S., Giraudeau, M., Vincze, O., Mcgraw, K. J., Aktipis, A., Tollis, M., & Maley, C. C. (2023). Life history and cancer in birds: clutch size predicts cancer. *BioRxiv*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9948971/>
- Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell,

- B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S. J., Frost, S. D. W., & Muse, S. V. (2020). HyPhy 2.5 - A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*, 37(1), 295–299. <https://doi.org/10.1093/molbev/msz197>
- Liu, Z., Shi, M., Ren, Y., Xu, H., Weng, S., Ning, W., Ge, X., Liu, L., Guo, C., Duo, M., Li, L., Li, J., & Han, X. (2023). Recent advances and applications of CRISPR-Cas9 in cancer immunotherapy. *Molecular Cancer*, 22(1), 1–19. <https://doi.org/10.1186/s12943-023-01738-6>
- Mitra, P. (2018). Transcription regulation of MYB: a potential and novel therapeutic target in cancer. *Annals of Translational Medicine*, 6(22), 443–443. <https://doi.org/10.21037/atm.2018.09.62>
- Møller, A. P., Erritzøe, J., & Soler, J. J. (2017). Life history, immunity, Peto's paradox and tumours in birds. *Journal of Evolutionary Biology*, 30(5), 960–967. <https://doi.org/10.1111/jeb.13060>
- Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, 8(7). <https://doi.org/10.1371/journal.pgen.1002764>
- Nature Cancer. (2020). The global challenge of cancer. *Nature Cancer*, 1(1), 1–2. <https://doi.org/10.1038/s43018-019-0023-9>
- Nenclares, P., & Harrington, K. J. (2020). The biology of cancer. *Medicine (United Kingdom)*, 48(2), 67–72. <https://doi.org/10.1016/j.mpmed.2019.11.001>
- Pullar, T. (1996). Kea (*Nestor notabilis*) Captive Management Plan and Husbandry Manual. *Threatened Species Occasional Publication*, 21(9), 1–15. <http://www.doc.govt.nz/Documents/science-and-technical/TSOP09.pdf>
- Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1), W191–W198. <https://doi.org/10.1093/nar/gkz369>
- Ruales, J., & Checa, F. (2018). La atención del cáncer en el Ecuador: pasado, presente y futuro. *Revista de La Facultad de Ciencias Médicas (Quito)*, 43(1), 46–59. https://doi.org/10.29166/ciencias_medicas.v43i1.1456
- Seim, I., Fang, X., Xiong, Z., Lobanov, A. V., Huang, Z., Ma, S., Feng, Y., Turanov, A. A., Zhu, Y., Lenz, T. L., Gerashchenko, M. V., Fan, D., Hee Yim, S., Yao, X., Jordan, D., Xiong, Y., Ma, Y., Lyapunov, A. N., Chen, G., ... Gladyshev, V. N. (2013). Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. *Nature Communications*, 4, 1–8. <https://doi.org/10.1038/ncomms3212>
- Seluanov, A., Gladyshev, V. N., Vijg, J., & Gorbunova, V. (2018). Mechanisms of cancer resistance in long-lived mammals. *Nature Reviews Cancer*, 18(7), 433–441. <https://doi.org/10.1038/s41568-018-0004-9>
- Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond,

- S. L. (2015). Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, 32(5), 1342–1353. <https://doi.org/10.1093/molbev/msv022>
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stiller, J., & Zhang, G. (2019). Comparative phylogenomics, a stepping stone for bird biodiversity studies. *Diversity*, 11(7), 1–19. <https://doi.org/10.3390/D11070115>
- Strasser, A., Jost, P. J., & Nagata, S. (2009). The Many Roles of FAS Receptor Signaling in the Immune System. *Immunity*, 30(2), 180–192. <https://doi.org/10.1016/j.immuni.2009.01.001>
- Tejada-Martinez, D., De Magalhães, J. P., & Opazo, J. C. (2021). Positive selection and gene duplications in tumour suppressor genes reveal clues about how cetaceans resist cancer. *Proceedings of the Royal Society B: Biological Sciences*, 288(1945). <https://doi.org/10.1098/rspb.2020.2592>
- Ulhas Nair, N., Cheng, K., Naddaf, L., Sharon, E., Pal, L. R., Rajagopal, P. S., Unterman, I., Aldape, K., Hannenhalli, S., Day, C. P., Tabach, Y., & Ruppin, E. (2022). Cross-species identification of cancer resistance-associated genes that may mediate human cancer risk. *Science Advances*, 8(31), 1–13. <https://doi.org/10.1126/sciadv.abj7176>
- Vincze, O., Colchero, F., Lemaître, J. F., Conde, D. A., Pavard, S., Bieuville, M., Urrutia, A. O., Ujvari, B., Boddy, A. M., Maley, C. C., Thomas, F., & Giraudeau, M. (2022). Cancer risk across mammals. *Nature*, 601(7892), 263–267. <https://doi.org/10.1038/s41586-021-04224-5>
- Wang, Q., Goldstein, M., Alexander, P., Wakeman, T. P., Sun, T., Feng, J., Lou, Z., Kastan, M. B., & Wang, X. F. (2014). Rad17 recruits the MRE11-RAD50-NBS1 complex to regulate the cellular response to DNA double-strand breaks. *EMBO Journal*, 33(8), 862–877. <https://doi.org/10.1002/embj.201386064>
- Wilson, K., Shiuan, E., & Brantley-Sieders, D. M. (2021). Oncogenic functions and therapeutic targeting of EphA2 in cancer. *Oncogene*, 40(14), 2483–2495. <https://doi.org/10.1038/s41388-021-01714-8>
- Yagmur Erten, E., Tollis, M., & Kokko, H. (2020). Bird size with dinosaur-level cancer defences: can evolutionary lags during miniaturisation explain cancer robustness in birds? *BioRxiv*, 2020.10.22.345439. <https://doi.org/10.1101/2020.10.22.345439>
- Zahn-Zabal, M., Dessimoz, C., & Glover, N. M. (2020). Identifying orthologs with OMA: A primer. *F1000Research*, 9. <https://doi.org/10.12688/f1000research.21508.1>

ANEXOS

ANEXO 1

<https://colab.research.google.com/drive/1PptrUdj4vk9Z-2XhW05GQNYE9Fb0b6DL?usp=sharing>

SCRIPT: DESCIFRANDO_CANCER_AVES

El siguiente pipeline permitirá descargar y analizar secuencias de genes codificantes de aves Psittaciformes ortólogas, con la finalidad de identificar posibles mutaciones únicas que se encuentren bajo selección positiva y potencialmente confieran resistencia al cáncer a este grupo de aves. Para ello, se realizarán los siguientes pasos:

1. Se descargarán las secuencias del NCBI. En este script se considera el protooncogén SRC(tirosina quinasa no receptora), para evaluar otros genes se debe modificar la línea de comandos de descarga.
2. Se realizará el alineamiento de secuencias con MUSCLE
3. Se generarán árboles filogenéticos con RAxML
4. Se evaluarán los resultados con HyPhy para identificar los sitios (MEME) y ramas (aBSREL)bajo selección.

Nota: Todo el script fue desarrollado en la terminal de Linux, en el editor de texto de línea de comando nano.

```
#!/bin/bash
```

DESCARGANDO SECUENCIAS DEL NCBI

1. Ingresar a la carpeta donde se descargarán los archivos con el comando `cd nombre_directorio`
2. Llamar al script de conda

```
source /home/manager/anaconda3/etc/profile.d/conda.sh
```

3. Se debe activar el conda environment de `ncbi_datasets`

```
conda activate ncbi_datasets
```

4. El siguiente comando se usa para descargar las secuencias:

- Se descarga según gene symbol, regiones codificantes y ortholog group

```
datasets download gene symbol src --include cds --ortholog Psittaciformes --filenam
```

5. Descomprimir documento

```
unzip practiceSRC.zip -d SRC-cds-practice
```

6. Se observa en árbol las carpetas donde está guardado el documento descomprimido

```
tree SRC-cds-practice
```

7. Desactivar el conda environment

```
conda deactivate
```

8. Se eliminan los archivos que no sean del formato .fna

```
find /home/manager/Data/SRC-cds-practice/ncbi_dataset/data -type f ! -name *.fna -
```

ALINEAMIENTO DE SECUENCIAS CON MUSCLE

```
muscle -in SRC-cds-practice/ncbi_dataset/data/cds.fna -out /home/manager/Data/SRC-cds-practice/ncbi_d
```

GENERANDO ÁRBOLES FILOGENÉTICOS CON RAXML

Primero se debe realizar modificaciones en el documento generado por muscle

1. Para editar el encabezado de cada secuencia del archivo y eliminar todo desde ":" utilizar el siguiente comando

```
cd /home/manager/Data/SRC-cds-practice/ncbi_dataset/data  
awk '/^>/{ sub(/:.*$/, "", $0) } 1' SRCmuscleAlignPrac.afa > SRCmuscleAlignMod1.afa
```

2. Para remover los stop codons utilizamos la función cln (clean) de HyPhy

```
hyphy cIn Universal /home/manager/Data/SRC-cds-practice/ncbi\_dataset/data/SRCmuscle
```

3. OPCIONAL: visualizar el archivo modificado (las primeras 40 líneas)

```
cat SRCmuscleAlignMod2.afa|head -n 40
```

4. Llamar al script de conda

```
source /home/manager/anaconda3/etc/profile.d/conda.sh
```

5. Se debe activar el conda environment donde se encuentra RAxML

```
conda activate phylogenetics
```

6. Generar el árbol filogenético (tree)

```
raxmlHPC -s /home/manager/Data/SRC-cds-practice/ncbi\_dataset/data/SRCmuscleAlignMod
```

7. Desactivar el conda environment

```
conda deactivate
```

8. Se eliminan los archivos que no sean necesarios, ej. los archivos que empiezan con el patrón RAxML_bipartitions* y RAxML_bootstrap*. Se utilizará el archivo RAxML_bestTree

```
find /home/manager/Data/SRC-cds-practice/ncbi\_dataset/data -type f \( -name "RAxML_
```

9. Visualizar el árbol filogenético previamente descargado con raxml utilizando itoL (<https://itol.embl.de/>)

UTILIZANDO HyPhy

Tomar en cuenta los siguientes parámetros para usar la línea de comando para hyphy:

- ♦ LIBPATH=ruta del directorio: define el directorio donde se encuentran los archivos de la biblioteca HyPhy
- ♦ Especificar en alignment la ruta del codon data file (obtenido del alineamiento de muscle)
- ♦ Indicar en tree la ruta del árbol filogenético (obtenido con Raxml)
- ♦ En output especificar la ruta donde se guardarán los resultados (usar la extensión .json)

Para analizar el método MEME:

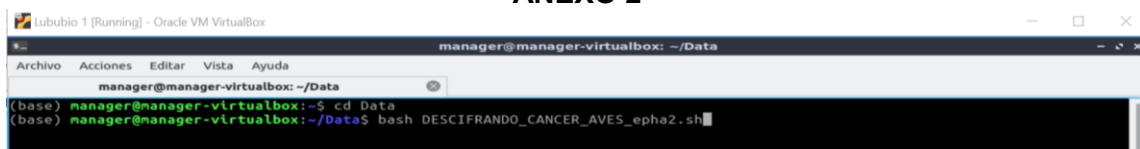
```
hyphy LIBPATH=/home/manager/anaconda3/lib/hyphy meme --alignment /home/manager/Data/SRC-cds-practice/
```

Para analizar el método aBSREL:

```
hyphy LIBPATH=/home/manager/anaconda3/lib/hyphy absrel --alignment /home/manager/Data/SRC-cds-practic
```

Visualizar los resultados de hyphy en formato de tabla cargando el documento .json en hyphyvision (<http://vision.hyphy.org>)

ANEXO 2



Descarga de secuencias y alineamiento con MUSCLE


```
manager@manager-virtualbox: ~/Data
Archivo Acciones Editar Vista Ayuda
manager@manager-virtualbox: ~/Data
This is RAXML version 8.2.12 released by Alexandros Stamatakis on May 2018.
With greatly appreciated code contributions by:
Andre Aberer (HITS)
Simon Berger (HITS)
Alexey Kozlov (HITS)
Kassian Kobert (HITS)
David Dao (KIT and HITS)
Sarah Lutteropp (KIT and HITS)
Nick Pattengale (Sandia)
Wayne Pfeiffer (SDSC)
Akifumi S. Tanabe (NRIFS)
Charlie Taylor (UF)
Alignment has 109 distinct alignment patterns
Proportion of gaps and completely undetermined characters in this alignment: 13.12%
RAXML rapid bootstrapping and subsequent ML search
Using 1 distinct models/data partitions with joint branch length optimization
Executing 1000 rapid bootstrap inferences and thereafter a thorough ML search
All free model parameters will be estimated by RAXML
ML estimate of 25 per site rate categories I
Likelihood of final tree will be evaluated and optimized under GAMMA
GAMMA Model parameters will be estimated up to an accuracy of 0.1000000000 Log Likelihood units
Partition: 0
Alignment Patterns: 109
Name: No Name Provided
DataType: DNA
Substitution Matrix: GTR
```

```
RAXML was called as follows:
raxmlHPC -s /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/EPHA2muscleAlignMod2.afa -n EPHA2raxml_trees -m GTRCAT -f a
-x 123 -p 256 -N autoMRE
Time for BS model parameter optimization 0.071868
Bootstrap[0]: Time 0.076604 seconds, bootstrap likelihood -4506.817404, best rearrangement setting 5
Bootstrap[1]: Time 0.001695 seconds, bootstrap likelihood -4513.127116, best rearrangement setting 6
Bootstrap[2]: Time 0.001950 seconds, bootstrap likelihood -4529.020412, best rearrangement setting 7
Bootstrap[3]: Time 0.002399 seconds, bootstrap likelihood -4496.709291, best rearrangement setting 11
Bootstrap[4]: Time 0.001832 seconds, bootstrap likelihood -4548.083226, best rearrangement setting 15
Bootstrap[5]: Time 0.001431 seconds, bootstrap likelihood -4502.091111, best rearrangement setting 15
Bootstrap[6]: Time 0.001518 seconds, bootstrap likelihood -4472.421843, best rearrangement setting 5
Bootstrap[7]: Time 0.001265 seconds, bootstrap likelihood -4445.000855, best rearrangement setting 12
Bootstrap[8]: Time 0.001280 seconds, bootstrap likelihood -4504.600836, best rearrangement setting 11
Bootstrap[9]: Time 0.001076 seconds, bootstrap likelihood -4440.005440, best rearrangement setting 5
Bootstrap[10]: Time 0.001228 seconds, bootstrap likelihood -4483.332888, best rearrangement setting 13
Bootstrap[11]: Time 0.001422 seconds, bootstrap likelihood -4535.765751, best rearrangement setting 7
Bootstrap[12]: Time 0.001366 seconds, bootstrap likelihood -4519.318960, best rearrangement setting 6
Bootstrap[13]: Time 0.001447 seconds, bootstrap likelihood -4554.274808, best rearrangement setting 5
Bootstrap[14]: Time 0.001445 seconds, bootstrap likelihood -4541.568493, best rearrangement setting 5
```

```
ML search took 0.502717 secs or 0.000140 hours
Combined Bootstrap and ML search took 0.680358 secs or 0.000189 hours
Drawing Bootstrap Support Values on best-scoring ML tree ...
Found 1 tree in File /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/RAXML_bestTree.EPHA2raxml_trees
Found 1 tree in File /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/RAXML_bestTree.EPHA2raxml_trees
Program execution info written to /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/RAXML_info.EPHA2raxml_trees
All 50 bootstrapped trees written to: /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/RAXML_bootstrap.EPHA2raxml_trees
Best-scoring ML tree written to: /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/RAXML_bestTree.EPHA2raxml_trees
Best-scoring ML tree with support values written to: /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/RAXML_bipartitions.EPHA2raxml_trees
Best-scoring ML tree with support values as branch labels written to: /home/manager/Data/EPHA2-cds-practice/ncbi_dataset/data/RAXML_bipartitionsBranchLabels.EPHA2raxml_trees
Overall execution time for full ML analysis: 0.681263 secs or 0.000189 hours or 0.000008 days
```

HyPhy aBSREL

```

manager@manager-virtualbox: ~/Data
Analysis Description
-----
HyPhy (Adaptive branch-site random effects likelihood) uses an adaptive
random effects branch-site model framework to test whether each branch
has evolved under positive selection, using a procedure which infers an
optimal number of rate categories per branch.
--Requirements--: in-frame codon alignment and a phylogenetic tree
--Citation--: Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient
Detection of Episodic Diversifying Selection (2015). Mol Biol Evol 32
(5): 1342-1353
--Written by--: Sergei L Kosakovsky Pond, Ben Murrell, Steven Weaver and Temple GEM /
UCSD viral evolution group
--Contact Information--: spond@temple.edu
--Analysis Version--: 2.1

--code --> Universal
-Loaded a multiple sequence alignment with **4** sequences, **1001** codons, and **1** partitions from '/home/manager/Data/EPHA
2-cds-practice/ncbi_dataset/data/EPHA2muscleAlignMod2.afa'
--branches --> All

### * Selected 5 branches for testing: 'XM_030507884_1, XM_057401309_1, XM_034068124_1, Node2, XM_010012765_1'

### Obtaining branch lengths and nucleotide substitution biases under the nucleotide GTR model
* Log(L) = -5120.59, AIC-c = 10267.20 (13 estimated parameters)

### Fitting the baseline model with a single dN/dS class per branch, and no site-to-site variation.
* Log(L) = -4854.20, AIC-c = 9756.69 (24 estimated parameters)
* branch-level non-synonymous/synonymous rate ratio distribution has median 4.44, and 95% of the weight in 10000000000.00 - 9
.67

### Determining the optimal number of rate classes per branch using a step up procedure


| Branch         | Length | Rates | Max. dN/dS     | Log(L)   | AIC-c   | Best AIC-c so f |
|----------------|--------|-------|----------------|----------|---------|-----------------|
| Node2          | 0.03   | 2     | 8.96 (49.95%)  | -4853.95 | 9760.25 | 9756.69         |
| XM_034068124_1 | 0.02   | 2     | 14.89 (63.13%) | -4854.17 | 9760.70 | 9756.69         |
| XM_030507884_1 | 0.01   | 2     | 45.44 (90.68%) | -4854.41 | 9761.18 | 9756.69         |
| XM_057401309_1 | 0.01   | 2     | 4.98 (100.00%) | -4854.20 | 9760.74 | 9756.69         |
| XM_010012765_1 | 0.01   | 2     | 37.21 ( 8.38%) | -4852.44 | 9757.24 | 9756.69         |



### Rate class analyses summary
* 5 branches with **1** rate classes
### Improving parameter estimates of the adaptive rate class model
* Log(L) = -4845.70, AIC-c = 9739.70 (24 estimated parameters)

### Testing selected branches for selection


| Branch         | Rates | Max. dN/dS      | Test LRT | Uncorrected p-value |
|----------------|-------|-----------------|----------|---------------------|
| Node2          | 1     | 5.44 (100.00%)  | 20.72    | 0.00001             |
| XM_034068124_1 | 1     | 11.87 (100.00%) | 28.73    | 0.00000             |
| XM_030507884_1 | 1     | >1000 (100.00%) | 20.68    | 0.00001             |
| XM_057401309_1 | 1     | 6.03 (100.00%)  | 10.19    | 0.00212             |
| XM_010012765_1 | 1     | 9.51 (100.00%)  | 3.37     | 0.06893             |



### Adaptive branch site random effects likelihood test
Likelihood ratio test for episodic diversifying positive selection at Holm-Bonferroni corrected  $p = 0.0500$ , found **4** bran
ches under selection among **5** tested.
* XM_034068124_1, p-value = 0.00000
* Node2, p-value = 0.00004
* XM_030507884_1, p-value = 0.00003
* XM_057401309_1, p-value = 0.00423

Check messages.log for diagnostic messages.
(base) manager@manager-virtualbox:~/Data$

```

HyPhy MEME

```

Lububio 1 [Running] - Oracle VM VirtualBox
manager@manager-virtualbox: ~/Data
--Contact Information--: spond@temple.edu
--Analysis Version--: 2.1.1

--code --> Universal
-Loaded a multiple sequence alignment with **4** sequences, **1001** codons, and **1** partitions from '/home/manager/Data/EPHA
2-cds-practice/ncbi_dataset/data/EPHA2muscleAlignMod2.afa'
--branches --> All
-Select the p-value threshold to use when testing for selection (permissible range = [0,1], default value = 0.1):
pvalue -> 0.1

### Branches to include in the MEME analysis
Selected 5 branches to include in the MEME analysis: 'XM_030507884_1, XM_057401309_1, XM_034068124_1, Node2, XM_010012765_1'

### Obtaining branch lengths and nucleotide substitution biases under the nucleotide GTR model
* Log(L) = -5120.59, AIC-c = 10267.20 (13 estimated parameters)

### Obtaining the global omega estimate based on relative GTR branch lengths and nucleotide substitution biases
* Log(L) = -4856.56, AIC-c = 9753.34 (20 estimated parameters)
* non-synonymous/synonymous rate ratio for *test* = 6.6428

### Improving branch lengths, nucleotide substitution biases, and global dN/dS ratios under a full codon model
* Log(L) = -4848.12
* non-synonymous/synonymous rate ratio for *test* = 7.5603

### For partition 1 these sites are significant at  $p \leq 0.1$ 

| Codon | Partition | alpha | beta*    | p*    | LRT   | Episodic selection detected? | # branches |
|-------|-----------|-------|----------|-------|-------|------------------------------|------------|
| 38    | 1         | 0.000 | 1770.302 | 0.332 | 3.170 | Yes, p = 0.0973              | 0          |
| 572   | 1         | 0.000 | 464.586  | 0.245 | 3.656 | Yes, p = 0.0757              | 1          |
| 597   | 1         | 0.000 | 1804.474 | 0.205 | 8.488 | Yes, p = 0.0063              | 1          |


```