UNIVERSIDADE DA CORUÑA
Facultad de Ciencias

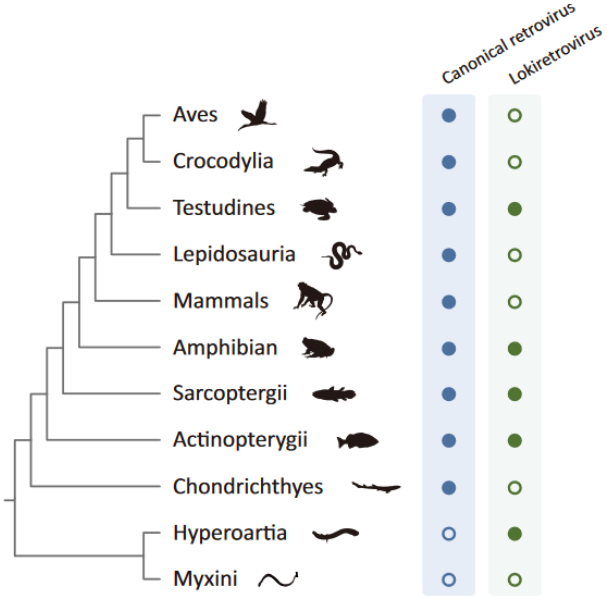
Grado en Biología

Memoria del Trabajo de Fin de Grado

Relaciones filogenéticas de los Lokiortervirus presentes en los genomas de los Petromyzontiformes

Relacións filoxenéticas dos Lokiortervirus presentes nos xenomas dos Petromyzontiformes

Phylogenetic relationships of the Lokiortervirus present in the genomes of Petromyzontiformes



Lucía Sánchez Abad

Curso: 2021- 2022. Convocatoria: Junio

Director: Horacio Naveira Fachal

Índice

Resumen

Desde hace unos 500 Ma, los Petromyzontiformes (lampreas) evolucionan de forma independiente con respecto al resto de los vertebrados, representando uno de los linajes más basales que se conocen. En anteriores trabajos, se describió la presencia de un grupo muy peculiar de retrovirus endógenos en el genoma de *Petromyzon marinus* y de *Lethenteron camtschaticum*, inicialmente denominados lokiretrovirus, y más recientemente lokiortervirus, que cuenta con representantes en un gran número de especies de peces, anfibios y reptiles, por lo que su origen podría ser, en principio, muy antiguo. Con todo, no se sabe con seguridad si esto es cierto o si su origen es mucho más reciente y su extensa distribución es debida a transmisión horizontal. En consecuencia, en este trabajo se propuso estudiar la presencia de representantes de esta clase de retrovirus endógenos en los genomas actualmente secuenciados de los Petromyzontiformes (*Petromyzon marinus, Entosphenus tridentatus, Lethenteron reissneri* y *Lethenteron camtschaticum*) y las relaciones filogenéticas que se establecen entre ellos. Además, se pretendía determinar la abundancia de provirus y la existencia de incongruencias entre los árboles filogenéticos de los retrovirus y de las especies hospedadoras, así como entre los obtenidos a partir de distintos dominios retrovirales. Para llevar a cabo este estudio, se realizó un análisis *in silico* utilizando bases de datos publicadas en el NCBI, que fueron minadas para localizar las inserciones presentes en el genoma de los Petromyzontiformes. Una vez caracterizadas, se utilizaron para construir árboles filogenéticos. Los resultados demuestran que los lokiortervirus están “vivos” dentro del género *Lethenteron* y que se transponen activamente a lo largo de su genoma. Esto podría indicar que la inserción de estos retrovirus en el genoma de las lampreas se habría producido antes de lo previsto o que aparecen en el genoma de este género por transmisión horizontal. Por otro lado, se observó la existencia de dos episodios diferentes de endogenización de los lokiortervirus. También los resultados demuestran la presencia de transmisión horizontal por mezcla de inserciones pertenecientes a distintas especies a lo largo de los distintos clados. Por último, se hallaron incoherencias entre las filogenias de los dominios RH y RT, indicativas de una evolución en mosaico.

Palabras clave: retrovirus endógenos, ERVs, lokiretrovirus, transmisión horizontal, recombinación, RNAsa H, retrotranscriptasa.

Resumo

Dende hai 500 Ma, os Petromyzontiformes (lampreas) evolucionan de forma independente con respecto ó resto dos vertebrados, representando unha das liñaxes máis basais coñecidas. En anteriores traballos, describiuse a presenza dun grupo moi peculiar de retrovirus endóxenos no xenoma de *Petromyzon marinus* e de *Lethenteron camtschaticum*, inicialmente denominados lokiretrovirus, e máis recentemente lokiortervirus, que conta con representantes nun gran número de especies de peixes, anfibios e réptiles, polo que a súa orixe podería ser, en principio, moi antiga. Con todo, non se sabe con seguridade se isto é certo ou se a súa orixe é moito máis recente e a súa extensa distribución é debida a transmisión horizontal. En consecuencia, neste traballo propúxose estudar a presenza de representantes desta clase de retrovirus endóxenos nos xenomas actualmente secuenciados dos Petromyzontiformes (*Petromyzon marinus*, *Entosphenus tridentatus*, *Lethenteron reissneri* e *Lethenteron camtschaticum*) e as relacións filoxenéticas que se establecen entre eles. Ademais, pretendíase determinar a abundancia de provirus e a existencia de incongruencias entre as árbores filoxenéticas dos retrovirus e das especies hospedadoras, así como entre as

obtidas a partir de distintos dominios retrovirais. Para levar a cabo este estudo, realizouse unha análise *in silico* empregando bases de datos publicadas no NCBI, que foron minadas para localizar as inserciónss presentes no xenoma dos Petromyzontiformes. Unha vez caracterizadas, foron empregadas para construír árbores filoxenéticas. Os resultados demostran que os lokiortervirus están “vivos” dentro do xénero *Lethenteron* e que se traspoñen activamente ó longo do seu xenoma. Isto podería indicar que a inserción destes retrovirus no xenoma das lampreas tivo lugar antes do previsto ou que aparecen no xenoma deste xénero por transmisión horizontal. Por outro lado, observouse a existencia de dous episodios diferentes de endoxenización dos lokiortervirus. Tamén, os resultados amosan a presenza de transmisión horizontal por mestura de inserciónss pertencentes a distintas especies ao longo dos distintos clados. Por último, atopáronse incoherencias entre as filoxenias dos dominios RH e RT, indicativas dunha evolución en mosaico.

Palabras claves: retrovirus endóxenos, ERVs, lokiretrovirus, transmisión horizontal, recombinación, RNAsa H, retrotranscriptasa.

Abstract

Over the last 500 Myr, Petromyzontiformes (lampreys) have evolved independently from the rest of vertebrates, representing one of the most basal lineages currently known. Previous work has described the presence of a very peculiar group of endogenous retroviruses in the genomes of *Petromyzon marinus* and *Lethenteron camtschaticum*, initially called lokiretroviruses, and more recently lokiorterviruses, which have representatives in a large number of fish, amphibian and reptile species, so that their origin could, in principle, be very ancient. However, it is uncertain whether this is true or whether its origin is much more recent and its widespread distribution is due to horizontal transmission. Consequently, this work aimed to study the presence of representatives of this class of endogenous retroviruses in the currently sequenced genomes of Petromyzontiformes (*Petromyzon marinus*, *Entosphenus tridentatus*, *Lethenteron reissneri* and *Lethenteron camtschaticum*) and the phylogenetic relationships between them. In addition, the aim was to determine the abundance of proviruses and the existence of inconsistencies between the phylogenetic trees of retroviruses and host species, as well as between those obtained from different retroviral domains.To carry out this study, an *in silico* analysis was performed using databases published in the NCBI, which were screened to locate the insertions present in the genome of Petromyzontiformes. Once characterised, they were used to create phylogenetic trees. The results show that lokiorterviruses are "alive" within the genus *Lethenteron* and that they are actively transposed throughout their genome. This could indicate that the insertion of these retroviruses into the lamprey genome would have occurred earlier than expected or that they appear in the genome of this genus by horizontal transmission. On the other hand, the existence of two different episodes of endogenisation of lokiorterviruses was observed. The results also demonstrate the presence of horizontal transmission by mixing insertions belonging to different species across clades. Finally, inconsistencies were found between the phylogenies of RH and RT domains indicating a mosaic evolution.

Key words: endogenus retrovirus, ERVs, lokiretrovirus, horizontal transmission, recombination, RNAse H, retrotranscriptase.

1. Introducción

Existen cinco familias de virus que se retrotranscriben. Estas están extendidas a lo largo de los genomas de animales, plantas, algas y hongos. Debido a sus características comunes, hoy en día, todas ellas están encuadradas dentro del orden Ortervirales (Krupovic et al., 2018).

Es frecuente que los miembros de este orden sean, comúnmente, denominados retrovirus. Sin embargo, los retrovirus son, propiamente, tan sólo los pertenecientes a la familia *Retroviridae*, restringida a especies de vertebrados. Estos provocan una gran variedad de enfermedades, entre ellas el Síndrome de Inmunodeficiencia Adquirida (SIDA) y cáncer. En este último caso, el cáncer es provocado a través de tres mecanismos: por transducción de oncogenes del hospedador, al actuar como mutágenos de inserción o a través de la actividad tumoral de los productos génicos virales (Coffin et al., 1997; Goff, 2007; Telesnitsky, 2010; Zheng et al., 2022).

Son virus formados por una partícula proteica que contiene dos copias de ssRNA de entre 7 y 10 kb de longitud. En general, los genomas de todos los retrovirus presentan tres genes: *Gag, Pol* y *Env* (Figura 1). El gen *Gag* codifica las proteínas de la matriz (MA), la capside (CA) y la nucleocapside (NC). Por otro lado, el gen *Pol* codifica la proteasa (PR), la retrotranscriptasa (RT), la RNasa H (RH) y la integrasa (IN). Finalmente, el gen *Env* codifica las glucoproteínas tansmembrana (TM) y de la superficie (SU) (Coffin et al., 1997; Goff, 2007; Telesnitsky, 2010; Zheng et al., 2022).

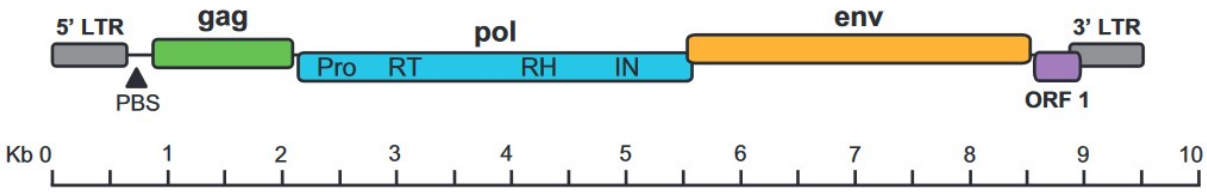


Figura 1. Estructura típica de la copia proviral del genoma de los retrovirus. El genoma de un provirus típico está formado por los genes *Gag, Pol* y *Env* flaqueados por Repeticiones Largas Terminales (LTRs). El gen *Pol* contiene los dominios de la proteasa (Pro), la retrotranscriptasa (RT), la ribonucleasa H (RH) y la integrasa (IN).

Nota. Adaptado de Wei et al. (2019).

Para la replicación del genoma de los retrovirus es necesaria la intervención de la retrotranscriptasa. Esta transcribe el RNA viral a dsDNA y lo integra en el genoma del hospedador como provirus. En una infección retroviral típica, esta inserción se produciría en las células somáticas del hospedador. Sin embargo, en ocasiones, infectan células de la línea germinal al insertarse en el genoma. De esta forma, el genoma vírico flanqueado por dos Repeticiones Terminales Largas (LTRs) es heredado verticalmente a la siguiente generación como parte de los cromosomas del hospedador. Así, se forma lo que se conoce como retrovirus endógenos (ERVs). Tanto en las inserciones en la línea germinal como en la somática, el DNA provírico se transcribe a partir del promotor situado en el LTR5'. El RNA se traduce formando las partículas proteicas infectivas donde se engloba el RNA vírico y que continuarán con la infección. (Goff, 2007; Stoye, 2012; Johnson, 2015, 2019; Zheng et al., 2022).

Al carecer la retrotranscriptasa de habilidad correctora, los retrovirus presentan tasas elevadas de mutación. Sin embargo, los ERVs evolucionan de forma conjunta con sus hospedadores en forma de loci genómicos. Este hecho supone una tasa de evolución mucho más baja (propia de regiones pseudogénicas) (Duffy et al., 2008; Gago et al., 2009; Zheng et al., 2022).

Por otro lado, entre todos estos genes, se encuentra un doble dominio de Ribonucleasa H (RH). La presencia de este doble dominio es debido a que, en el momento en el que el virus adquiere un nuevo dominio RH, el anterior degenera formando lo que se conoce como dominio Tether (Malik & Eickbush, 2001; Smyshlyaev et al., 2013; Ustyantsev et al., 2015; Wang & Han, 2021).

En cuanto, a su clasificación, los retrovirus exógenos se pueden encuadrar dentro de dos familias: *Orthoretrovirinae* y *Spumaretroviridae*. En la primera de ellas, se pueden observar seis géneros: *Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus y Lentivirus*. Por otro lado, la segunda presenta cinco géneros: *Bovispumavirus, Equispumavirus, Felispumavirus, Prosimiispumavirus y Simiispumavirus* (Walker et al., 2019; Wang & Han, 2021). En cambio, los ERVs pueden agruparse en las clases I (relacionada con los *Gammaretrovirus* y los *Epsilonretrovirus*), II (relacionada con los *Betaretrovirus*) y III (relacionada con los *Spumavirus*) (R. Gifford & Tristem, 2003; R. J. Gifford et al., 2018; Wang & Han, 2021). Con todo, estas dos clasificaciones no son compatibles debido a que el sistema de clasificación de los ERVs no ha sido bien diseñado y presenta numerosos problemas prácticos (R. J. Gifford et al., 2018; Xu et al., 2018; Wang & Han, 2021).

En un principio, se sitúa el origen de los retrovirus en la evolución temprana de los vertebrados, antes del origen de los vertebrados con mandíbula, hace 450 Ma. La distribución de los ERVs en vertebrados, la ausencia de ERVs en el genoma de lanceta (*Branchistoma floridae*) (perteneciente al subfilo Cephalocordata, cercanamente relacionado con el subfilo Vertebrata) y su ausencia fuera de los vertebrados apoya este hecho (Hedges et al., 2015; Hayward, 2017; Xu et al., 2018; Zheng et al., 2022).Con todo, la distribución de los ERVs podría ser explicada a través de un origen más tardío y una propagación en los vertebrados a través de frecuentes intercambios de hospedadores.

La primera teoría está apoyada por el hecho de que los spumavirus (una subfamilia de retrovirus) co-divergen con sus hospedadores vertebrados mandibulados durante más de 450 Ma (Han & Worobey, 2012; Aiewsakun & Katzourakis, 2017; Xu et al., 2018; Wei et al., 2019; Chen et al., 2021; Zheng et al., 2022). Además, los retrovirus hallados en tetrápodos encajan dentro de la diversidad de retrovirus de peces. Estos ocupan posiciones basales en los clados de retrovirus más importantes. Teniendo todo esto en cuenta, los retrovirus se habrían originado en el medio acuático hace más de 450 Ma (Xu et al., 2018; Zheng et al., 2022).

En cualquier caso, durante mucho tiempo, se creyó que este origen habría tenido lugar a partir de un retrotransposón antiguo Ty3/Gypsy (Doolittle et al., 1989; Xiong & Eickbush, 1990; Hayward, 2017; Wang & Han, 2022). A pesar de ello, los únicos retrotransposones muestreados pertenecientes a esta clase estaban relacionados de forma lejana con los retrovirus. En consecuencia, no fue hasta el descubrimiento de los retrotranspones Odin, un linaje descendiente moderno de dichos retrotransposones, que se dio credibilidad a dicho origen (Wang & Han, 2022).

Diversos análisis filogenómicos indican que estos retrovirus surgidos hace más de 450 Ma aparecen en el genoma de todos los vertebrados mandibulados estudiados (Hayward et al., 2013; Xu et al., 2018; Zheng et al., 2021, 2022). Por otro lado, si reparamos en los vertebrados no mandibulados, se han identificado ERVs en la lamprea *Petromyzon marinus*. No sólo es un hallazgo particular por la presencia de estos retrovirus en vertebrados no mandibulados, sino porque más tarde serían renombrados en un nuevo grupo. Inicialmente, este sería denominado lokiretrovirus. Con todo, el hallazgo del linaje de retrotransposones Odin en los genomas de 8 anémonas de mar (orden Actinaria)

dentro del filo Cnidaria lo harían ser renombrado como lokiortervirus (Hayward et al., 2015; Xu et al., 2018; Wang & Han, 2021, 2022).

Los lokiortervirus están presentes en el genoma de vertebrados, como las lampreas, los peces de aleta lobuladas, los de aletas con radios, los anfibios y los reptiles (Johnson, 2015; Wang & Han, 2021; Zheng et al., 2022). Según Wang y Han (2021), este nuevo linaje podría representar una nueva subfamilia contenida dentro de la familia *Retroviridae*. Así pues, basándose en los análisis filogenéticos del dominio RT, formarían un grupo hermano con respecto al resto de retrovirus (Johnson, 2015; Wang & Han, 2021; Zheng et al., 2022).

Sin embargo, la consideración de los retrotransposones Odin como grupo hermano de los lokiortervirus ha provocado, a su vez, a un cambio en la relación de estos últimos con los retrovirus canónicos (*Orthoretrovirinae* y *Spumaretroviridae*). De esta manera, ya no son considerados un grupo monofilético, sino que podrían representar dos familias de virus distintas (Wang & Han, 2022).

Según los trabajos de Wang y Han (2021), las inserciones provirales más antiguas de lokiortervirus datan de hace aproximadamente 37 Ma (como es el caso de la especie *Nanorana parkeri*), y aparentemente estos elementos siguen transponiéndose activamente en algunas especies, o lo han hecho hasta un pasado muy reciente.

Como se puede ver en la Figura 2, los lokiortervirus presentan una estructura del genoma similar a la del resto de retrovirus: 3 ORFs (*gag, pol* y *env*) flanqueados por LTRs. A pesar de ello, presentan alguna diferencia con respecto a los otros retrovirus que los hacen únicos (Wang & Han, 2021).

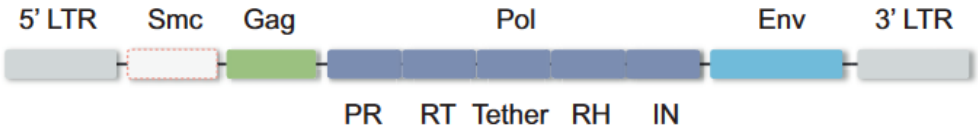


Figura 2. Estructura del genoma consenso del provirus de lokiortervirus. El genoma del provirus de lokiortervirus codifica al menos 3 ORFs (*gag, pol* y *env*) flanqueados por dos Repeticiones Largas Terminales (LTRs). El genoma de algunos provirus codifica un gen adicional llamado *Smc.*

Nota. Adaptado de Wang & Han (2021).

Una de ellas, como se puede observar en la Figura 2, es la presencia del gen *Smc*, ausente en otros retrovirus. Este gen codifica una proteína que es homóloga de las que se encargan de mantener la estructura de los cromosomas. Este tipo de proteínas se unen al DNA implicándose en las dinámicas de los cromosomas (Harvey et al., 2002). Según Wang y Han (2021), se habría adquirido a lo largo de la evolución de los lokiortervirus.

Por otro lado, las secuencias de las proteínas de ENV no comparten similitudes con los retrovirus. Sin embargo, sí comparten similitudes con las glucoproteínas de ciertos virus del orden Mononegavirales. Esto indicaría que ambas proteínas derivan del mismo ancestro viral común (Wang & Han, 2021). En consecuencia, el gen *Env* de retrovirus y el de lokiortervirus tienen orígenes distintos. Este hecho junto con la afirmación mencionada con anterioridad de que no conforman un grupo monofilético es lo que da pie al renombramiento del grupo de los lokiretrovirus en lokiortervirus (Wang & Han, 2022).

En cuanto a los dominios RH e IN, diversos análisis filogenéticos han evidenciado la presencia de múltiples grupos distintivos. Este hecho se explica a través del continuo reemplazamiento de estos dominios durante su evolución. De esta manera, la complejidad

del genoma de los retrovirus podría explicarse a través de este barajado de dominios (Wang & Han, 2021).

En el caso del dominio RH, el dominio presente en lokiortervirus y spumavirus es considerado el representante del linaje RH más antiguo de los retrovirus. Como tanto los lokiortervirus como los retrovirus presentan un dominio dual formado por un dominio RH degenerado (Tether) y un dominio RH recién adquirido, se considera que la degradación del dominio RH preexistente habría tenido lugar con anterioridad a la aparición del ancestro común más reciente de estos dos linajes. Este domino Tether de los retrovirus es muy similar al de los metavirus y, probablemente, habría sido adquirido a partir de un hospedador eucariota. Lo mismo sucedería con el nuevo dominio adquirido por los retrovirus tras la divergencia de los spumavirus (Wang & Han, 2021).

Si se profundiza un poco más en los animales en los que se han encontrado inserciones de lokiortervirus, los trabajos de Xu et al. (2018) destacan, a escala del genoma, la presencia de inserciones en la lamprea de mar (*P. marinus*) y su ausencia en la lamprea ártica (*Lethenteron camtschaticum*).

Se estima que los lokiortervirus invadieron los genomas de las lampreas hace unos 27-34 Ma (Xu et al., 2018), lo que podría explicar la presencia de representantes de este grupo tanto en *P. marinus* como en *L. camtschaticum*, ya que ambas especies iniciaron su divergencia hace unos 30-38 Ma (Kuraku & Kuratani, 2006).

Por otro lado, las inserciones halladas en *P. marinus* son cercanas a las de peces de aletas con radios y de aletas lobuladas. En consecuencia, los ERVs presentes en la lamprea de mar no se corresponderían con un linaje retroviral antiguo. En cambio, este hecho puede sugerir una transmisión entre especies más recientes (Hedges et al., 2015; Xu et al., 2018).

Otro hecho a destacar son las conclusiones que se generan al comparar la filogenia de los lokiortervirus construida en base al dominio RT con la filogenia de los vertebrados. Un ejemplo de ello son *Anguilla rostrata* y las lampreas. Las inserciones de lokiortervirus presentes en *A. rostrata* divergen antes que las presentes en las lampreas. En cambio, en el árbol filogenético de los hospedadores, las lampreas se separan antes que *A. rostrata* (Wang & Han, 2021).

Por otro lado, los análisis filogenéticos de Xu et al. (2018) muestran que la mayoría de los retrovirus no evidencian la relación filogenética cercana que se establece entre los hospedadores y los retrovirus de grupos vertebrados distintos. Por ejemplo, los ERVs presentes en peces cartilaginosos no ocupan posiciones basales dentro de ningún clado principal de retrovirus. Por el contrario, están distribuidos a lo largo del árbol filogenético.

Estos dos hechos evidencian los continuos y complejos cambios de hospedadores sufridos por los retrovirus a lo largo de su evolución. Es decir, durante su historia evolutiva es común que los retrovirus se transmitan horizontalmente (Xu et al., 2018; Wang & Han, 2021).

A pesar de todo esto, es más probable que la transmisión tenga lugar entre linajes cuyos nichos ecológicos son solapantes. Por ejemplo, todos los compañeros de transmisión de los peces con aletas con radios viven en ambientes acuáticos (Xu et al., 2018).

Además, es probable que las transmisiones que tienen lugar en los nudos terminales sólo reflejen aquellas más recientes. De esta manera, podrían estar teniendo

lugar transmisiones entre clases por medio de hospedadores intermedios. Así pues, se puede subestimar el número de eventos de transmisión entre clases (Xu et al., 2018).

Como hemos ido desgranando a lo largo de este apartado, en estos momentos se desconoce si las familias de retrovirus halladas en los Petromyzontiformes (lampreas) se corresponden con grupos supervivientes de los ya existentes cuando las lampreas iniciaron su divergencia del resto de los vertebrados hace 500 Ma (Smith et al., 2013) (Figura 3) o si su origen es más reciente siendo el resultado de procesos de transmisión horizontal desde alguna otra especie de pez. En consecuencia, este trabajo busca explorar las relaciones filogenéticas existentes entre los distintos representantes en los genomas secuenciados de los Petromyzontiformes: *Petromyzon marinus, Entosphenus tridentatus, Lethenteron reissneri* y *Lethenteron camtschaticum*. Los últimos datos sobre la filogenia de los Petromyzontiformes señalan que la primera especie en separarse del tronco común fue *P. marinus*, seguida más tarde por *E. tridentatus* y, por último, el linaje del que descienden *L. camtschaticum* y *L. reissneri* (ver Figura 2 de Pereira et al. (2021)). A través de esta exploración se podrían obtener las evidencias necesarias para resolver esta cuestión.

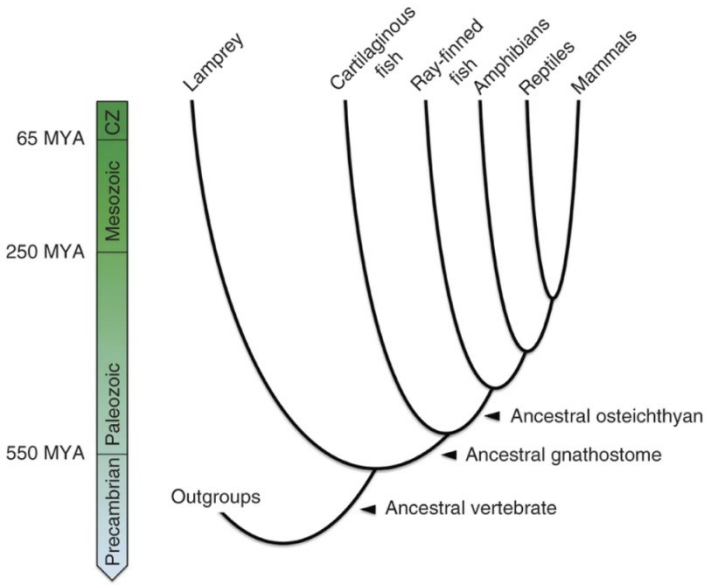


Figura 3. Filogenia de los vertebrados. Evidencia el momento de separación de las lampreas con respecto al resto de los vertebrados. Las lampreas ocupan una posición basal en la filogenia de los vertebrados.

Nota. Adaptado de Smith et al. (2013).

2. Objetivos

✓ Localizar y caracterizar las inserciones de ERVs pertenecientes a los lokiortervirus y presentes en el genoma de *Petromyzon marinus, Entosphenus tridentatus*, *Lethenteron reissneri* y *Lethenteron camtschaticum*.

✓ Estudiar las relaciones filogenéticas entre las inserciones en base al dominio retrotranscriptasa (RT) y el dominio Ribonucleasa H (RH), teniendo en cuenta tanto el dominio Tether como el nuevo dominio RH.

✓ Comparar los árboles filogenéticos de los distintos dominios en busca de incongruencias entre ellos. Estas serían indicativas de la existencia de recombinación (evolución en mosaico).

✓ Buscar indicios de transmisión horizontal a través de la presencia de incongruencias entre el árbol de especies y el árbol de ERVs.

3. Material y Métodos

3.1. Elaboración de la Query

El primer paso consistió en la elaboración de una query para localizar inserciones de lokiortervirus en el genoma de Petromyzontiformes. La secuencia de *Petromyzon marinus* empleada para obtener la query fue recuperada a partir del ERV1_Pet_mar_JAAIYE01000319_a del TFG de Rodríguez-Varela (2021).

Esta secuencia fue llevada al programa BioEdit (Hall, 1999). Una vez allí, se copió en formato FASTA para realizar una búsqueda en el apartado de "Conserved domains" de la página del National Center for Biotecnology Information (NCBI) (Lu et al., 2020). De esta forma, se pudieron localizar los dominios presentes en la secuencia y necesarios para construir las queries.

Se elaboraron dos queries, una que contenía el domino de la retrotranscriptasa (RT) y otra con el dominio de la RNAsa H (RH). En ambos casos, se anotaron las coordenadas del lugar del genoma donde se encontraba el dominio mostradas por el resultado de "Conserved domains" (Lu et al., 2020) (Figura E1).

A continuación, se buscaron dichas coordenadas en la secuencia que estaba abierta en BioEdit (Hall, 1999). Una vez acotadas, se copió ese fragmento y se creó una nueva secuencia con él dando lugar a la query. Esta secuencia se guardó con el nombre del dominio.

3.2. Localización y Caracterización de las Inserciones

El siguiente paso consistió en la localización de inserciones de lokiortervirus en el genoma de las lampreas de estudio (*Petromyzon marinus*, *Entosphenus tridentatus*, *Lethenteron reissneri* y *Lethenteron camtschaticum*). Para ello, se realizaron dos BLASTn (a través de la página del NCBI (https://www.ncbi.nlm.nih.gov/)) sobre el genoma de estas especies utilizando las dos queries elaboradas en el apartado anterior.

Con esta finalidad, primero se buscó el genoma de la especie de estudio en cuestión en el apartado "genome" de la página del NCBI (https://www.ncbi.nlm.nih.gov/). Seguidamente, una vez hallado el genoma, como se puede ver en la Figura 4, se utilizó la opción que dice "BLAST against "el nombre de la especie" genome", en vez de la opción" BLAST genome". Esto es debido a que la segunda forma utiliza una base de datos diferente (RefSeqGenomic). Esta no contiene aún los genomas secuenciados de los Petromyzontiformes. En consecuencia, no se obtuvieron resultados al realizar un BLAST siguiendo ese camino.

Figura 4. Resultado generado por la página del NCBI al buscar en el apartado "genome" la especie *Lethenteron reissneri*. En la página que contiene el genoma de la especie buscada nos proporcionan dos opciones de BLAST la encuadrada en verde (que fue la utilizada en este trabajo) y la encuadrada en rojo.

En todos los casos, se realizó un megaBLAST (Morgulis et al., 2008) utilizando como query o bien el dominio RT o bien el dominio RH (dependiendo del caso) bajo las condiciones que se observan en la Figura 5.

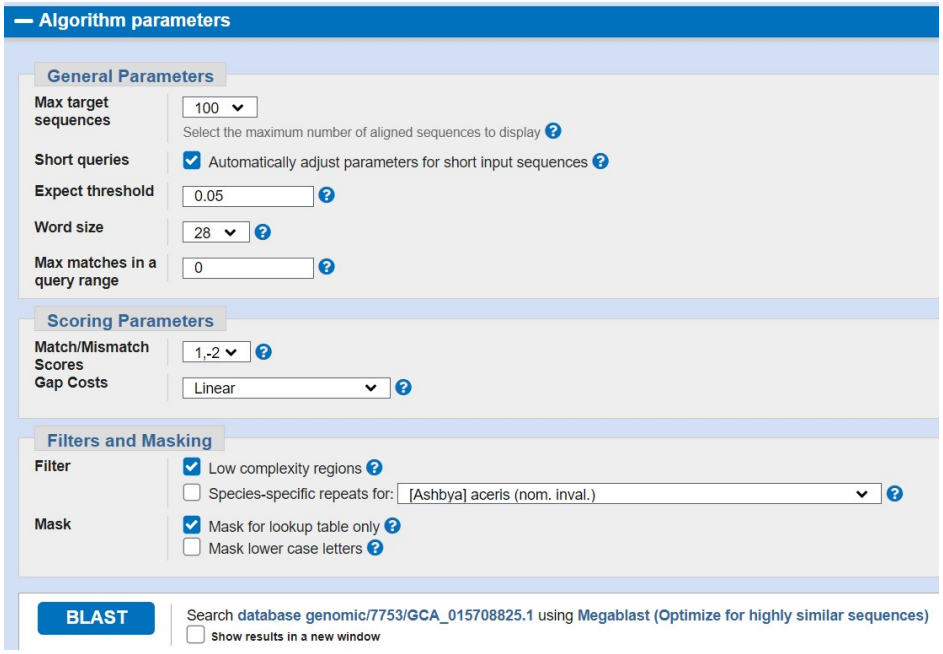


Figura 5. Parámetros empleados para realizar el megaBLAST.

De los hits proporcionados como resultados, se descartaron aquellos que dentro de su nombre presentaban la terminología "unplaced". Estas son regiones con muchas repeticiones que no han sido capaces de ensamblar y, por lo tanto, inútiles para este estudio.

En todos los casos, se descargaron todos los hits que presentaban el nivel máximo de ensamblaje. De manera que, si estaba disponible el nivel de cromosoma tanto el scaffold como los contigs eran descartados. En cambio, si este no lo estaba, sólo los contigs eran descartados.

Por otro lado, se observó la presencia de pares de hits iguales con diferentes "accession number" ("cromosomas duplicados"). En consecuencia, se descartó uno de los dos, eliminando siempre el mismo a lo largo de la especie.

Una vez obtenidos todos los hits, se descargó el Excel con todos los datos de los hits y las secuencias en formato FASTA de los escogidos. Para ello, se eligieron las opciones "Download", "Hit table (text)" y "FASTA (aligned sequence)". Las secuencias en formato FASTA se guardaron como "Hits-dominio-tres primeras letras de género y epíteto específico.fas".

Como los datos que deseábamos importar a formato Excel ("Hit table (text)") estaban en formato de texto, el primer paso consistió en abrir el programa Excel y

pasar los datos a dicho formato. A tal efecto, en la opción de “Datos” se clicó en la pestaña de “Obtener datos externos desde un archivo de texto” y se buscó el archivo guardado bajo el nombre de “Resultados BLAST.txt”. Finalmente, se guardó como “Resultados BLAST.xlsx”.

Una vez guardado el Excel con los resultados del BLAST, se procedió a modificarlo para adecuarlo a los intereses de esta investigación obteniendo el resultado que se describe y observa en el Anexo A.

3.3. Modificación de los Nombres de las Secuencias

El siguiente paso consistió en cambiar el nombre de las secuencias presentes en los archivos guardados bajo el nombre “Hits-dominio-tres primeras letras de género y epíteto específico.fas”. Todo esto, se llevó a cabo con la finalidad de obtener uno más corto e identificable a la hora de analizar las filogenias. Para facilitar la comprensión de dichos nombres, se elaboró un Excel con la correspondencia de esas nomenclaturas y la descripción de la secuencia que designan (Tabla B1).

Modificando los nombres se buscó transformar las largas designaciones proporcionados por el NCBI (https://www.ncbi.nlm.nih.gov/) (ej. "NC_046124.1:9144413-9144996 Petromyzon marinus isolate kPetMar1 chromosome 56, kPetMar1.pri") en un nombre que recogiese “tres primeras letras del género y tres primeras letras del epíteto específico_número de cromosoma/scaffold_coordenadas” (ej. petmar_56_9144413-9144996). Para tal fin, se utilizaron los programas RStudio (RStudio, 2021) y Excel en un proceso detallado en el Anexo B.

3.4. Archivo de Secuencias Sin Alinear

La siguiente tarea consistió en elaborar un archivo que contuviese todas las secuencias de las inserciones presentes en el genoma de las cuatro especies de estudio separadas por dominio.

En primer lugar, con ayuda del programa BioEdit (Hall, 1999) se juntaron todos los archivos “Hits-dominio-tres primeras letras de género y epíteto específico.fas” pertenecientes al mismo dominio en uno solo. Para ello, primero se abrió uno de los archivos de una especie y se seleccionaron todas las secuencias copiándolas al portapapeles. A continuación, se creó un nuevo archivo desde el portapapeles, conteniendo este todas las secuencias copiadas antes. A este nuevo archivo, se le añadieron el resto de las secuencias de las otras especies para el dominio en cuestión al copiarlas y pegarlas en él. Finalmente, se guardó bajo el nombre “erv1_dominio_all.fas”.

3.5. Alineamiento

El siguiente paso, consistió en el alineamiento de las secuencias presentes en el archivo “erv1_dominio_all.fas”. Para ello, tras seleccionar todas las secuencias presentes en el archivo, se ejecutó el programa Clustal W Multiple Alignment (Thompson et al., 1994) a través de BioEdit (Hall, 1999), manteniendo los parámetros por defecto. El archivo obtenido se guardó bajo el nombre “erv1_dominio_all_aln.fas”.

Por último, se revisó el alineamiento para buscar algún error en él y realizar las modificaciones adecuadas a través del proceso detallado en el Anexo C.

3.6. Creación de la Secuencia Consenso

El siguiente paso, consistió en la creación de una secuencia consenso con el objetivo de resolver las posibles dudas que se pudieran crear. Una secuencia consenso se crea empleando los nucleótidos que aparecen con más frecuencia en todas las secuencias utilizadas para su elaboración. Por lo tanto, idealmente se corresponde con la secuencia ancestral de la que derivan todas las demás.

Para crear la secuencia consenso, se utilizó de nuevo el programa BioEdit (Hall, 1999). Tras seleccionar todas las secuencias, simplemente se hizo uso de las opciones: Alignment>Create consensus sequence. Tras su creación, fue revisada y modificada a través de un proceso detallado en el Anexo D.

3.7. Separación del Dominio RH en Dominio Tether y Nuevo Dominio RH

Como ya se mencionó en la introducción, el dominio RH es un dominio dual formado por el domino Tether y el nuevo dominio RH, como se puede ver en la Figura E1. En consecuencia, es interesante separar en las secuencias ambos dominios. Teniendo esto en cuenta, se crearon dos nuevos archivos “erv1_dominio_all_aln.fas” con los dos nuevos dominios a través del proceso descrito en el Anexo E.

Con este último paso, ya están listos todos los archivos necesarios para la construcción de los árboles filogenéticos.

3.8. Creación de las Filogenias

El último paso consistió en la creación de los árboles filogenéticos. Se elaboraron 3 filogenias por cada uno de los archivos (“erv1_rt_all_aln.fas”, “erv1_rh_all_aln.fas”, “erv1_tether_all_aln.fas” y “erv1_new-rh_all_aln.fas”). Los métodos escogidos para elaborar las filogenias fueron:

- Método de Máxima Parsimonia bajo las condiciones observadas en la Figura 6A.
- Método de Neighbor-Joining bajo las condiciones observadas en la Figura 6B.
- Método de Máxima Verosimilitud bajo las condiciones de la Figura 6C.

En el caso de los dos primeros métodos, se utilizó el programa MEGA 11 (Tamura et al., 2021). En cambio, para el tercero se empleó el servidor IQTREE (Nguyen et al., 2015) debido a los problemas que daba MEGA11 (Tamura et al., 2021) a la hora de analizar nuestros datos a través del método de Máxima Verosimilitud.

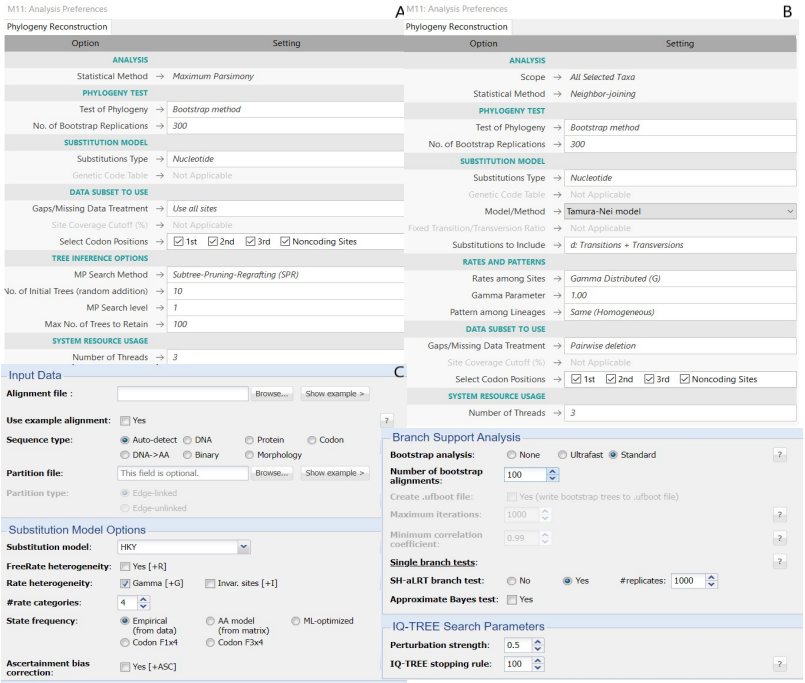


Figura 6. Condiciones de creación de las reconstrucciones filogenéticas. (A) Condiciones para el método de Máxima Parsimonia. (B) Condiciones para el método de Neighbor-Joining. (C) Condiciones para el método de Máxima Verosimilitud.

En el caso de los dos últimos métodos, fue necesario realizar un test para elegir un modelo de sustitución (Models>Find Best DNA/Proteins Models (ML)) bajo las condiciones observadas en la Figura 7. El modelo más adecuado será el que aparezca primero. Este será el que tenga el mayor BIC o bondad de ajuste del modelo estadístico. En los casos en los que no estaba disponible el mejor modelo, se empleó el segundo. A mayores, este test permitió determinar la opción “rates among sites” observable en los dos últimos modelos. De esta manera, las condiciones observadas en las Figuras 6B y 6C varían según el resultado de este test.

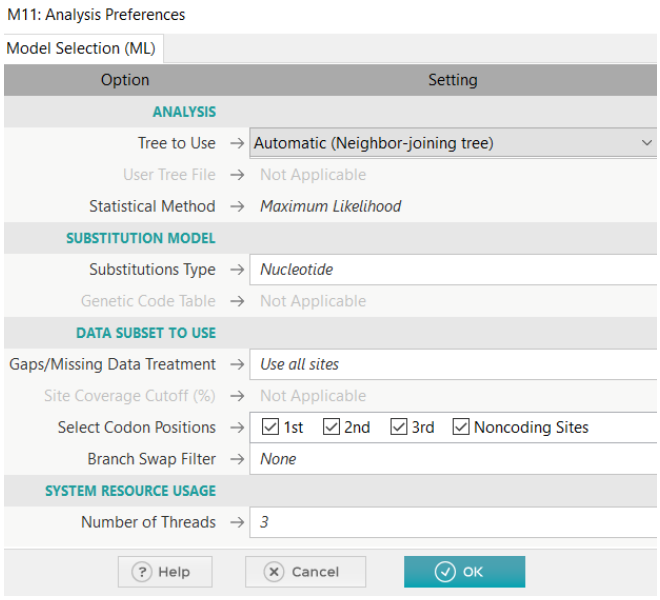


Figura 7. Condiciones del test para determinar el modelo a utilizar en los métodos de reconstrucción filogenética de Máxima Verosimilitud y Neighbor-joining.

En cuanto al enraizamiento de los árboles filogenéticos, se utilizó la rama que parte de las dos secuencias “petmar_68”. Estas probablemente no sean dos inserciones independientes, sino que una de ellas es el resultado de la duplicación de una región cromosómica que contenía a la inserción original. Estas dos, sin duda, representan la inserción más antigua de la colección de secuencias. Por lo tanto, son adecuados para ser utilizadas como outgroup.

¿Cómo se supo que eran las secuencias más antiguas y adecuadas para ser empleadas como outgroup?

Al construir un árbol filogenético con el método Neighbor-Joining (bajo las mismas condiciones de la Figura 6C), sin enraizar y con una disposición radial, se pudo observar una rama mucho más larga que destacaba entre las demás (Figura F1). Esta rama de mayor longitud se corresponde con la más antigua y, por lo tanto, adecuada para realizar el enraizamiento en base a ella.

Esta conclusión se apoyó con el conocimiento de que, cuando un retrovirus se inserta en el genoma como provirus, sus dos LTRs son idénticas. Con todo, a medida que transcurre el tiempo, van divergiendo por acumulación de mutaciones a un ritmo similar al de un pseudogen. En consecuencia, se puede estimar la antigüedad de una inserción a partir de las diferencias entre sus LTRs, si se conoce la tasa de evolución nucleotídica del DNA silencioso. Teniendo esto en cuenta, se calculó el porcentaje de identidad de las LTRs, obteniendo como resultado un 84%. Este implica una antigüedad de 26 Ma, aplicando una tasa de evolución neutra de $2.2 * 10^{-9}$ sustituciones/lugar/año (Wang & Han (2021)).

Una vez enraizados todos los árboles, se condensaron bajo un valor de cutoff del 85%. Seguidamente, se emplearon los árboles construidos bajo el método de Máxima Versomilitud como base para elaborar un árbol que contuviese los apoyos estadísticos para cada nudo conseguidos a través de los tres métodos empleados. En algunos casos, se observaron clados que aparecían en Máxima Parsimonia y Neighbor-Joining, pero que estaban ausentes en Máxima Verosimilitud. En consecuencia, se disminuyó el cutoff de dicha filogenia hasta que dichos clados apareciesen con un apoyo bootstrap.

4. Resultados

En este trabajo, se hizo un estudio de los dominios conservados presentes en el ERV1_Pet_mar_JAAIYE01000319_a recuperado del TFG de Rodríguez-Varela (2021). Al situar los dominios RT y RH en dicha secuencia, cuya posición se puede ver en la Figura E1, se pudieron localizar las inserciones presentes a lo largo del genoma de *Petromyzon marinus*, *Lethenteron reissneri*, *Lethenteron camtschaticum* y *Entosphenus tridentatus*.

De esta forma, en *P. marinus* (lamprea marina), utilizando como query el dominio RT, se hallaron inserciones de ERVs en 10 cromosomas diferentes de un total de 85 cromosomas (Tabla A1). En algunos casos, se encontró más de 1 inserción por cromosoma (como mucho 3 inserciones por cromosoma). Así pues, en total, se observaron 15 inserciones a lo largo de su genoma. También, se observaron dos casos en los que la presencia de indels reduce la longitud del alineamiento (inicio tardío del alineamiento) (cromosoma 49 y 68 (1074825-1075267)). Por último, en cuanto al porcentaje de identidad, las dos inserciones presentes en el cromosoma 68 presentan el mayor porcentaje de identidad (100% y 98.42%).

Por otro lado, utilizando como query el dominio RH, de nuevo, se encontraron inserciones de ERVs en 10 cromosomas diferentes de un total de 85 cromosomas (Tabla A5). En algunos casos, también se halló más de 1 inserción por cromosoma (como máximo 3 inserciones por cromosoma). De esta forma, en total, se observaron 15 inserciones a lo largo de su genoma, todas ellas localizadas a continuación del dominio RT siendo esta su posición natural como se puede observar en la Figura E1. A pesar de ello, como se puede ver en la Tabla A5, se observaron una serie de inserciones fragmentadas presentes en el dominio RH, pero ausentes en el dominio RT. También, se observaron una serie de casos donde, de nuevo, los indels reducen la longitud de los alineamientos. Al igual que en el dominio RH, las dos inserciones presentes en el cromosoma 68 presentan el mayor porcentaje de identidad (100% y 97.624%).

En cambio, en *L. reissneri* (lamprea de arroyo de Extremo Oriente), empleando como query el dominio RT, se hallaron inserciones de ERVs en 44 cromosomas diferentes de un total de 72 (Tabla A2). De nuevo, se observó más de 1 inserción por cromosoma en algunos casos (como máximo 9 provirus por cromosoma). En consecuencia, en total, se encontraron 99 inserciones. Nuevamente, se observó un caso en el que el inicio del alineamiento se produjo tardíamente (cromosoma 65 (4474250-4474807)). Por último, en cuanto al mayor porcentaje de identidad, este ya no está representado por las dos inserciones presentes en el cromosoma 68, sino por la inserción presente en el cromosoma 65 entre las coordenadas 4474250-4474807 (85.816%).

Por el contrario, utilizando como query el dominio RH, se encontraron inserciones en 39 cromosomas diferentes de un total de 72 (Tabla A6). Nuevamente, se halló más de 1 inserción por cromosoma en algunos casos (como máximo 9 por cromosoma). Por lo tanto, se observaron en total 90 inserciones, todas ellas a continuación del dominio RT. Con todo, se observaron algunas inserciones fragmentadas, tanto presentes en el dominio RT y ausentes en el dominio RH, como a la inversa, como se puede ver en la Tabla A6. En esta ocasión, una vez más, se observaron casos en los que el inicio del alineamiento se produjo tardíamente. Dicho esto, en algunos de ellos, aunque el programa BLAST consideraba esos primeros nucleótidos muy diferentes, en realidad no era así. Dentro de los que sí se cumple el inicio tardío del alineamiento, hemos de destacar la inserción en el cromosoma 14 situada entre las coordenadas 5133632-5133903. En este caso, el alineamiento comienza en la posición 65. En consecuencia, como gracias al proceso detallado en el Anexo E sabemos que el dominio Tether abarca de las posiciones 1-268, esta inserción en el cromosoma 14 no presentaría dicho dominio y el nuevo dominio RH empezaría muy tardíamente. Finalmente, al igual que en el dominio RT, la inserción presente en el cromosoma 65 entre las coordenadas 4475077-4476003 posee el mayor porcentaje de identidad (91.909 %).

Por otro lado, en *L. camtschaticum* (lamprea ártica), empleando como query el dominio RT, se hallaron inserciones en 27 scaffolds diferentes (Tabla A3). Al igual que en las especies anteriores, en algunos casos, se encontró más de 1 inserción por scaffold (como máximo 5 por scaffold). Así pues, se observaron un total de 39 inserciones. De nuevo, se encontró algún caso en el que el inicio del alineamiento se produjo tardíamente (inserción en el scaffold 53 entre las coordenadas 5644652-5645210). Finalmente, a diferencia de *P. marinus* y *L. reissneri*, la inserción que presenta el mayor porcentaje de identidad se encuentra en el scaffold 53 entre las coordenadas 5644652-5645210 (86.17%).

En cambio, al utilizar como query el dominio RH, se encontraron inserciones en 24 scaffolds diferentes (Tabla A7). Nuevamente, se observó más de 1 inserción por scaffold (como máximo 5 por scaffold). De manera que, en total, se hallaron 34 inserciones, todas

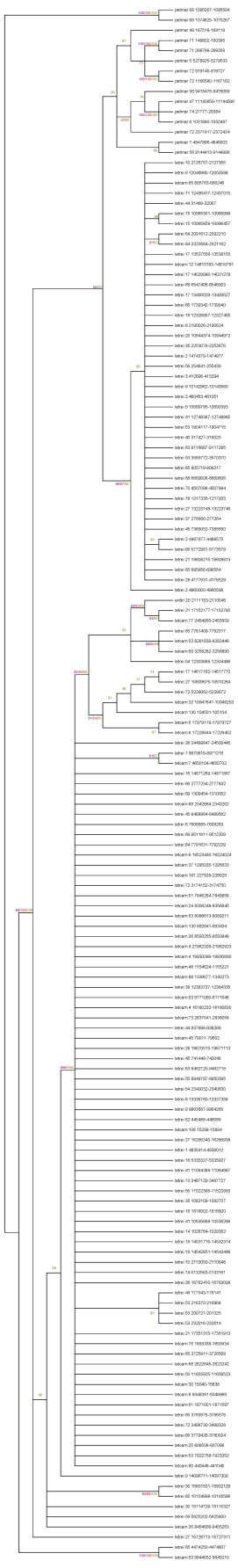


Figura 8. Árbol filogenético elaborado en base al archivo de datos “erv1_rt_all_aln.fas”.

ellas situadas a continuación del dominio RT. A pesar de ello, como se puede ver en la Tabla A7, se observaron algunas inserciones fragmentadas, tanto presentes en el dominio RT y ausentes en el dominio RH como a la inversa. Por otra parte, al igual que en las especies anteriores, se observaron casos en los que el inicio del alineamiento se produjo tardíamente. En todos ellos, aunque el programa BLAST consideraba esos primeros nucleótidos muy diferentes, en realidad no era así (salvo en los que solo se retrasaba el inicio del alineamiento 1 nucleótido). Por último, al igual que en el dominio RT, es la inserción presente en el scaffold 53 entre las coordenadas 5643465-5644387 la que presenta el mayor porcentaje de identidad (92.565 %).

Finalmente, en *E. tridentatus* (lamprea del Pacífico), tanto empleando como query el dominio RT (Tabla A4) como el dominio RH (Tabla A8), se observó una única inserción en el cromosoma 20.

Para cumplir con el segundo objetivo de este trabajo, se estudiaron los árboles filogenéticos correspondientes a los dominios RT y RH, dejando los formados por el dominio Tether y el nuevo dominio RH para trabajos a futuro. En primer lugar, se analizará la filogenia correspondiente al dominio RT reconstruida a partir de los métodos Neighbor-Joining (NJ), Máxima Parsimonia (MP) y Máxima Verosimilitud (ML) (Figura 8).

Se observan dos clados muy bien separados del resto y con alto apoyo bootstrap (UFBoot=100% para los 3 métodos). Uno de ellos está formado por las dos inserciones petmar_68 que se han usado para enraizar el árbol y el otro por las inserciones letrei_65_4474250-4474807 y letcam_53_5644652-5645210.

El otro clado bien diferenciado está también fuertemente apoyado (UFBoot=100% para Máxima Parsimonia y Máxima Verosimilitud y UFBoot=99% para Neighbor-Joining). En general, en este se observa como las inserciones pertenecientes a distintas especies se encuentran formando parte del mismo clado. Un ejemplo de esto sería el que se encuentra por encima del formado por letrei_65_4474250-4474807 y letcam_53_5644652-5645210. En él podemos ver tanto inserciones pertenecientes a *L. reissneri* como *L. camtschaticum* formando parte del mismo clado con un alto apoyo bootstrap.

Otro clado de relevancia es el que contiene a la única inserción de *E. tridentatus*. Este presenta un fuerte apoyo bootstrap (UFBoot=92% para NJ y MP y UFBoot=88%).

Figura 8. *Leyenda:* Esta filogenia se reconstruyó utilizando las inserciones pertenecientes al dominio RT presentes en el genoma de los Petromyzontiformes de estudio. Se elaboraron 3 árboles empleando los métodos NJ, MP y ML. Utilizando como base la filogenia del método de ML, se añadieron en los nodos correspondientes los apoyos estadísticos de los otros dos métodos. De esta forma, en cada nodo se observa el bootstrap obtenido con Neighboor-Joining, Máxima Parsimonia y Máxima Verosimilitud.

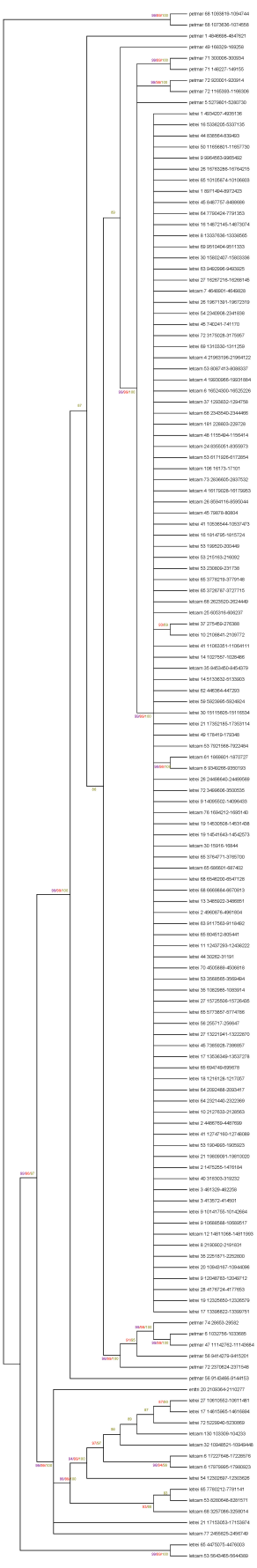


Figura 9. Árbol filogenético elaborado en base al archivo de datos “erv1_rh_all_aln.fas”.

En este clado, aparte de la inserción de la lamprea del Pacífico, se observan otras dos inserciones: una perteneciente a *L. reissneri* y la otra a *L. camtschaticum*. Además, este grupo se encuentra dentro de otro más numeroso compuesto por inserciones pertenecientes a estas dos especies.

Otro hecho a destacar, puede ser la presencia de un grupo de inserciones pertenecientes a *P. marinus* con un apoyo estadístico moderadamente elevado (UFBoot=90% para el método de Máxima Verosimilitud). Dentro de este, hay un grupo formado por cinco secuencias con apoyo estadístico significativo tanto para el método de Máxima Verosimilitud y Máxima Parsimonia (UFBoot=92% para ambos métodos).

A continuación, se analizará el árbol filogenético correspondiente al dominio RH (Figura 9).

Al igual que en el caso anterior, observamos dos clados muy separados del resto y con alto apoyo bootstrap (UFBoot=99% para NJ y MP y UFBoot=100% para ML). Estos están formados por los mismos provirus que en la filogenia del dominio RT.

Nuevamente, tenemos un segundo gran clado separado de estos otros dos con un gran apoyo bootstrap (UFBoot=99%, UFBoot=90% y UFBoot=97%). En él, también se observa de forma general como las inserciones de diferentes especies están mezcladas formando parte de los mismos clados. Un ejemplo de todo esto sería el que se encuentra en la parte superior del árbol y que empieza con la inserción letrei_1_4934207-4935136. Apoyado por un alto valor de bootstrap (UFBoot = 99% para Neighbor-Joining y Máxima Parsimonia y UFBoot=100% para Máxima Verosimilitud), se pueden ver tanto inserciones pertenecientes a *L. reissneri* como a *L. camtschaticum*.

Una vez más, otro clado de relevancia sería el que contiene a la única inserción de *E. tridentatus* apoyado por un alto valor de bootstrap (UFBoot=99% para Neighbor-Joining y Máxima Parsimonia y UFBoot=100% para Máxima Verosimilitud). Con todo, en contraposición al árbol filogenético del dominio RT, la localización de este clado es diferente. En esta ocasión, este clado se separa tempranamente del resto de miembros con los que comparte ancestro común.

Por último, al igual que en la filogenia anterior, se puede destacar un grupo de 5 inserciones de *P. marinus* con fuerte apoyo bootstrap para los 3 métodos (UFBoot=96%, UFBoot=99% y UFBoot=100%).

Figura 9. *Leyenda:* Esta filogenia se reconstruyó utilizando las inserciones pertenecientes al dominio RH presentes en el genoma de los Petromyzontiformes de estudio. Se elaboraron 3 árboles empleando los métodos NJ, MP y ML. Utilizando como base la filogenia del método de ML, se añadieron en los nodos correspondientes los apoyos estadísticos de los otros dos métodos. De esta forma, en cada nodo se observa el bootstrap obtenido con Neighboor-Joining, Máxima Parsimonia y Máxima Verosimilitud.

5. Discusión

En este estudio, se han hallado numerosos provirus a lo largo del genoma de *Petromyzon marinus, Lethenteron reissneri, Lethenteron camtschaticum* y *Entosphenus tridentatus.* En trabajos anteriores como los de Xu et al. (2018), ya se había determinado la presencia de ERVs en el genoma de *P. marinus* (lamprea marina). Sin embargo, se detallaba su ausencia en el genoma de *Lethenteron*, concretamente en el genoma de *L. camtschaticum* (lamprea ártica). Probablemente, esto fue debido a que, en el momento en el que se realizó dicho trabajo, su genoma aún no había sido secuenciado. No fue hasta el año 2021 cuando se secuenciaron los genomas de las especies *L. reissneri* y *L. camtschaticum*. En cambio, nuestro trabajo no sólo detalla la presencia de provirus en el genoma del género *Lethenteron,* sino que estos son mucho más abundantes que en el de *P. marinus* y mucho más parecidos entre sí. Todo esto parece indicar dos cosas:

Que los ERVs se están transponiendo muy activamente a lo largo del genoma de género *Lethenteron* (en mayor medida en el de *L. reissneri* ya que la abundancia de provirus es mayor). Por otro lado, se observa una menor actividad en *P. marinus* y una actividad prácticamente nula en *E. tridentatus*.

Que los lokiortervirus invadieron los genomas de las lampreas antes de la divergencia de la lamprea marina y la ártica al contrario de lo que avanzaba Xu et al. (2018) y, por lo tanto, de forma más temprana a lo estimado.

Por otro lado, tanto para el dominio RT como para el dominio RH, se observan dos clusters de provirus totalmente separados del resto (Figuras 8 y 9). Esto se explica debido a la existencia de dos episodios diferentes de endogenización de los lokiortervirus:

- La primera ola supondría la aparición de los provirus en el cromosoma 68 de *P. marinus*, letrei_65_4474250-4474807/ 4475075-4476003 y letcam_53_5644652-5645210/5643465-5644389. Esta es mucho más antigua y de menor éxito. Esto último se razona en base a que sólo se observa 1 copia en un sólo cromosoma de estas 3 especies.
- La segunda ola daría lugar al resto de las inserciones siendo mucho más reciente. Fue mucho más exitosa debido a una mayor capacidad de proliferación y transposición muy activa dando lugar a una gran cantidad de inserciones, como se puede ver en las Figuras 8 y 9.

  Esta alta capacidad proliferativa fue debida a que sus ancestros comunes (ramas más internas) estuvieron expuestos a una selección purificadora que permitió que las secuencias ancestrales conservaran los dominios necesarios para permanecer activas, a pesar de que una vez que se insertan en la línea germinal tienden a comportarse como pseudogenes (“dead on arrival”) (ramas más externas). La transición a pseudogen surge debido a las mutaciones que, a lo largo de las sucesivas generaciones, inactivan el provirus. Con todo, basta una única copia retroviral activa para continuar con la proliferación.

Una cuestión siempre presente al plantear la evolución de los elementos transponibles es la importancia que pueden haber jugado los eventos de transmisión horizontal en contraposición a la transmisión vertical. Estos, muy probablemente, sean facilitados, como ya mencionaban Gifford y Tristem (2003) y Xu et al. (2018), por el solapamiento de los nichos ecológicos o por características de la historia vital del hospedador. Si hablamos del solapamiento de los nichos ecológicos, todas las especies analizadas en este estudio son miembros de los Petromyzontiformes, que viven en un ambiente acuático (viven en el mar, pero desovan en agua dulce) y tienen los mismos hábitos de alimentación. Por lo tanto, el hecho de tener las mismas características y

desarrollar modos de vida similares podría provocar que la transmisión horizontal fuese muy frecuente. Esta cuestión se resuelve atendiendo a las discrepancias que puedan observarse entre la filogenia de los elementos transponibles y la filogenia de las especies.

Así pues, atendiendo al criterio anterior, podría decirse que las pocas inserciones producidas por la primera ola de endogenización son debidas a transmisión vertical. Esto es debido a que su filogenia es congruente con la de las especies indicada en la "Introducción". Esto, a su vez, implicaría que el provirus original estaba fijado en el ancestro común. En consecuencia, este fue transmitido a todos los descendientes del ancestro común de lampreas (transmisión vertical). Por lo tanto, dicha inserción se habría perdido en *E. tridentatus*.

Una forma de corroborar la hipotésis anterior sería realizar un estudio de las relaciones de sintenia existentes entre estas especies. Con todo, en un análisis preliminar, se ha podido concluir que las 3 regiones en donde se encuentran estas inserciones son ortólogas. Por lo tanto, estas tres inserciones se habrían originado por transmisión vertical. También, parece corresponder a un patrón de transmisión vertical la agrupación significativa de buena parte de las inserciones de *P. marinus*, que excluye a las de las otras especies.

Sin embargo, en la segunda ola, son abundantes los indicios de transmisión horizontal (incongruencias entre el árbol de especies y la filogenia de inserciones). El caso más llamativo es el de la única inserción encontrada en *E. tridentatus*, la cual se encuentra dentro de un clado que contiene otras inserciones de *L. reissneri* y *L. camtschaticum*. Esto llama mucho la atención ya que, si no hubiese transmisión horizontal, las inserciones de estas especies estarían totalmente separadas.

Por otro lado, al comparar los árboles filogéneticos de ambos dominios, como se menciona en el apartado de "Resultados", se observan incongruencias. Este tipo de inconsistencias se deben buscar en las relaciones más profundas, más cercanas a la raíz. Entre ellas, la más clara y evidente es la variación en la posición del clado que contiene a la única inserción de *E. tridentatus*. Este ejemplo, nos permite concluir que existe recombinación, la cual es signo de una evolución en mosaico.

Estas comparaciones entre ambas filogenias son hechas por el ojo humano y no tienen ningún apoyo estadístico. Por lo tanto, un objetivo para futuros trabajos consistiría en realizar un análisis de las diferencias topológicas entre ambas filogenias que permita obtener dicho apoyo estadístico. Una opción sería realizar un test topológico con ayuda del servido IQTREE (http://iqtree.cibiv.univie.ac.at/).

Finalmente, la presencia de más de un provirus por cromosoma puede indicar dos cosas:

A) Aparecen como resultado de la duplicación de la inserción original. Este es, muy probablemente, el caso de las dos inserciones presentes en el cromosoma 68 de *P. marinus*. Esto sería debido a que son muy parecidas entre sí (100% y 97.624% de identidad, ver Tabla A1 y A5). Además, como se puede ver con los porcentajes de identidad en las Tablas A1 y A5 y a través de las filogenias de las Figuras 8 y 9, son muy diferentes con respecto al resto y entre ambas existe poca distancia evolutiva.

B) La inserción del lokiortervirus se produce en el cromosoma más de una vez a través de eventos distintos e independientes.

Por consiguiente, un objetivo para futuros trabajos podría consistir en dictaminar si esta presencia de más de una inserción por cromosoma es debida a una duplicación o a una nueva inserción a través de un nuevo evento totalmente independiente del primero. Existen dos formas posibles de averiguarlo:

- Analizar el "Target Side Duplication" (TSD). En el caso de Petromyzontiformes, se produce la duplicación de los 4 nucleótidos anteriores al lugar de la inserción. De manera que, estos 4 nucleótidos se encontrarán antes de la LTR5' y después de la LTR3'. Si los nucleótidos antes y después de las LTRs son iguales, estaríamos ante una duplicación. En cambio, si son diferentes estaríamos ante eventos independientes de inserción.
- Realizar un BLAST 2 a 2 entre las inserciones. Para ello, se abriría un intervalo que contuviese al dominio (por ejemplo 10 Kb antes y después). De esta forma, si sólo se produce el alineamiento en la región del dominio, las inserciones serían el resultado de eventos independientes. Por el contrario, si el alineamiento se produce a lo largo de todo el intervalo, serían inserciones duplicadas.

Siguiendo con esta línea, se comentaba en el apartado de "Resultados" la presencia de inserciones fragmentadas (presentes en un dominio, pero ausentes en el otro). Esto podría indicar que, a lo largo de la evolución, los dominios RT en un caso y los dominios RH en el otro han sufrido mutaciones, como indels, inversiones, deficiencias, etc., o recombinaciones que han hecho desparecer o han alterado el dominio generando un provirus incompleto, truncado. Así mismo, esta explicación sería aplicable para aquellas inserciones donde se produce un inicio más tardío del alineamiento.

6. Conclusiones

6.1. Conclusiones

1) Los lokiortervirus están presentes en el genoma del género *Lethenteron*. Se observa una transposición activa de estos tanto en *L. reissneri* como en *L. camtschaticum*, con una menor intensidad en este último. En *Petromyzon marinus*, se observa una transposición mucho menos intensa que en los dos anteriores. Esta parece ser nula en *Entosphenus tridentatus*.

2) La presencia de los lokiortervirus en el genoma del género *Lethenteron* indica que su inserción se produjo antes de la divergencia de *P. marinus* y *L. camtschatium*.

3) A lo largo de la evolución, se produjeron dos episodios diferentes de endogenización de los lokiortervirus en el genoma de los Petromyzontiformes, siendo el segundo el de mayor éxito y, probablemente, todavía activo.

4) Las inserciones de la primera ola son el resultado de una transmisión vertical. En cambio, las de la segunda ola evidencian frecuentes eventos de transmisión horizontal.

5) La existencia de incongruencias entre los árboles filogenéticos de los dominios RT y RH evidencia la presencia de recombinación y, por lo tanto, de una evolución en mosaico.

6) La aparición de más de un provirus por cromosoma implica o una duplicación de la inserción original o la endogenización de más de un lokiortervirus a partir de eventos independientes.

7) La observación de inserciones fragmentadas y de indels que reducen la longitud de los alineamientos implica la existencia de provirus incompletos por mutaciones o recombinaciones durante su evolución.

8) El número de inserciones detalladas en este trabajo, probablemente, es una subestima debido a la fragmentación de sus secuencias. Esta dificulta su detección con los métodos utilizados.

9) En estudios futuros, se podría:

a. Comparar las filogenias realizadas en base a al dominio Tether y al nuevo dominio RH, con el objetivo de observar las incongruencias entre ambas. Estas podrían surgir debido al hecho de que el nuevo dominio RH es funcional y está sujeto a presiones selectivas. Además, estas también pueden ser debidas a la adquisición más tardía del nuevo dominio RH.

b. Analizar si la aparición de más de un provirus por cromosoma es debida a la duplicación de la inserción original o a eventos independientes.

c. Estudiar las relaciones de sintenia existentes entre el cromosoma 68 de *P. marinus*, el cromosoma 65 de *L. reissneri* y el scaffold 53 de *L. camtschaticum*.

d. Realizar tests estadísticos de la incongruencia topológica existente entre las filogenias obtenidas para los dominios RT y RH.

6.2. Conclusións

1) Os lokiortervirus están presentes no xenoma do xénero *Lethenteron.* Obsérvase unha transposición activa destes tanto en *L. reissneri* como en *L. camtschaticum*, cunha menor intensidade neste último. En *Petromyzon marinus*, obsérvase unha transposición moito menos intensa en comparación cos dous anteriores. Esta parece ser nula en *Entosphenus tridentatus*.

2) A presenza dos lokiortervirus no xenoma do xénero *Lethenteron* indica que a súa inserción tivo lugar antes da diverxencia de *P. marinus* e *L. camtschaticum*.

3) Ao longo da evolución, tiveron lugar dous episodios diferentes de endoxenización dos lokiortervirus no xenoma dos Petromyzontiformes, sendo o segundo o de maior éxito e, probablemente, aínda activo.

4) As insercións da primeira vaga son o resultado dunha transmisión vertical. En cambio, as da segunda onda evidencian frecuentes eventos de transmisión horizontal.

5) A existencia de incongruencias entre as árbores filoxenéticas dos dominios RT e RH evidencia a presenza de recombinación e, polo tanto, dunha evolución en mosaico.

6) A aparición de máis dun provirus por cromosoma implica ou ben unha duplicación da inserción orixinal ou ben a endoxenización de máis dun lokiortervirus a partir de eventos independentes.

7) A observación de insercións fragmentadas e de indels que reducen a lonxitude dos aliñamentos implica a existencia de provirus incompletos por mutacións ou recombinacións durante a súa evolución.

8) O número de insercións detalladas neste traballo, probablemente, é unha subestima debido á fragmentación das súas secuencias. Esta dificulta a súa detección a través dos métodos empregados.

9) En estudos futuros, poderíase:

a. Comparar as filoxenias realizadas en base ó dominio Tether e ó novo dominio RH, co obxectivo de observar as incoherencias entre ambas. Estas poderían xurdir debido ao feito de que o novo dominio RH é funcional e está suxeito a presións selectivas. Ademais, estas incongruidades tamén poderían ser debidas á adquisición máis tardía deste dominio.

b. Analizar se a aparición de máis dun provirus por cromosoma é debida á duplicación da inserción orixinal ou a eventos independentes.

c. Estudar as relacións de sintenia existentes entre o cromosoma 68 de *P. marinus*, o cromosoma 65 de *L. reissneri* e o scaffold 53 de *L. camtschaticum*.

d. Realizar tests estatísticos da incongruencia topolóxica existente entres as filoxenias obtidas para os dominios RT e RH.

6.3. Conclusions

1) Lokiortervirus “are alive” in the genome of the genus *Lethenteron*. They are actively transposing in *L. reissneri* and *L. camtschaticum*, with less intensity in the last one. In *Petromyzon marinus,* the transposition is much less active than in the other two and absent in *Entosphenus tridentatus*.

2) The presence of the lokiortervirus in the genome of the genus *Lethenteron* indicates that their insertion took place before the divergence between *P. marinus* and *L. reissneri*.

3) In the course of evolution, two different episodes of lokiortervirus endogenisation took place in the genome of Petromyzontiformes, being the second one the most successful and, probably, still active.

4) The insertions of the first wave are the result of vertical transmission. In contrast, the insertions of the second wave are evidence of frequent horizontal transmission events.

5) The existence of contradictions between the phylogenies of RT and RH domains evidence the presence of recombination and, therefore, of mosaic evolution.

6) The appearance of more than one provirus per chromosome indicates either a duplication of the original insertion or the creation of more than one ERV through independent events.

7) The observation of fragmented insertions and indels, which cause the reduction of the length of alignment, indicates the existence of incomplete provirus. The reason of these could be the presence of mutations and recombinations during their evolution.

8) The number of insertions reported in this work is probably an underestimation due to the fragmentation of their sequences. This makes difficult to identify them with the methods we have used.

9) Future objectives:

a. Compare the phylogenies of the Tether domain and the new RH domain with the goal of observing the contradictions between both of them. The reason of these discrepancies could be that the new RH domain is functional and because of that is under selective pressures. In addition, these may also be due to the later acquisition of the new RH domain.

b. Analyse the explanation of the presence of more than one provirus per chromosome (the duplication of the original insertion or an independent event of insertion).

c. Relationships of synteny between chromosome 68 of *P. marinus*, chromosome 65 of *L. reissneri* and scaffold 53 of *L. camtschaticum* could be studied.

d. Perform statistical tests of the topological incongruence between the phylogenies obtained for the RT and RH domains.

7. Bibliografía

Aiewsakun, P., & Katzourakis, A. (2017). Marine origin of retroviruses in the early Palaeozoic era. *Nature Communications*, *8*(1), 13954. https://doi.org/10.1038/ncomms13954

Chen, Y., Zhang, Y.-Y., Wei, X., & Cui, J. (2021). Multiple infiltration and cross-species transmission of foamy viruses across the Paleozoic to the Cenozoic era. *Journal of Virology*, *95*(14), e00484-21. https://doi.org/10.1128/JVI.00484-21

Coffin, J. M., Hughes, S. H., & Varmus, H. E. (1997). The interactions of retroviruses and their hosts. En J. M. Coffin, S. H. Hughes, & H. E. Varmus (Eds.), *Retroviruses*. Cold Spring Harbor Laboratory Press. https://www.ncbi.nlm.nih.gov/books/NBK19465

Doolittle, R. F., Feng, D.-F., Johnson, M. S., & McClure, M. A. (1989). Origins and evolutionary relationships of retroviruses. *The Quarterly Review of Biology*, *64*(1), 1-30. https://doi.org/10.1086/416128

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: Patterns and determinants. *Nature Reviews Genetics*, *9*(4), 267-276. https://doi.org/10.1038/nrg2323

Gago, S., Elena, S. F., Flores, R., & Sanjuán, R. (2009). Extremely high mutation rate of a hammerhead viroid. *Science*, *323*(5919), 1308. https://doi.org/10.1126/science.1169202

Gifford, R., & Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, *26*(3), 291-315. https://doi.org/10.1023/A:1024455415443

Gifford, R. J., Blomberg, J., Coffin, J. M., Fan, H., Heidmann, T., Mayer, J., Stoye, J., Tristem, M., & Johnson, W. E. (2018). Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology*, *15*, 59. https://doi.org/10.1186/s12977-018-0442-1

Goff, S. P. (2007). Host factors exploited by retroviruses. *Nature Reviews Microbiology*, *5*(4), 253-263. https://doi.org/10.1038/nrmicro1541

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95-98.

Han, G.-Z., & Worobey, M. (2012). Endogenous lentiviral elements in the weasel family (*Mustelidae*). *Molecular Biology and Evolution*, *29*(10), 2905-2908. https://doi.org/10.1093/molbev/mss126

Harvey, S. H., Krien, M. J. E., & O'Connell, M. J. (2002). Structural maintenance of chromosomes (SMC) proteins, a family of conserved ATPases. *Genome Biology*, *3*(2), reviews3003.1. https://doi.org/10.1186/gb-2002-3-2-reviews3003

Hayward, A. (2017). Origin of the retroviruses: When, where, and how? *Current Opinion in Virology*, *25*, 23-27. https://doi.org/10.1016/j.coviro.2017.06.006

Hayward, A., Cornwallis, C. K., & Jern, P. (2015). Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proceedings of the National Academy of Sciences*, *112*(2), 464-469. https://doi.org/10.1073/pnas.1414980112

Hayward, A., Grabherr, M., & Jern, P. (2013). Broad-scale phylogenomics provides insights into retrovirus–host evolution. *Proceedings of the National Academy of Sciences*, *110*(50), 20146-20151. https://doi.org/10.1073/pnas.1315419110

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, *32*(4), 835-845. https://doi.org/10.1093/molbev/msv037

Johnson, W. E. (2015). Endogenous retroviruses in the genomics era. *Annual Review of Virology*, *2*, 135-159. https://doi.org/10.1146/annurev-virology-100114-054945

Johnson, W. E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. *Nature Reviews Microbiology*, *17*(6), 355-370. https://doi.org/10.1038/s41579-019-0189-2

Krupovic, M., Blomberg, J., Coffin, J. M., Dasgupta, I., Fan, H., Geering, A. D., Gifford, R., Harrach, B., Hull, R., Johnson, W., Kreuze, J. F., Lindemann, D., Llorens, C., Lockhart, B., Mayer, J., Muller, E., Olszewski, N. E., Pappu, H. R., Pooggin, M. M., … Kuhn, J. H. (2018). *Ortervirales*: New virus order unifying five families of reverse-transcribing viruses. *Journal of Virology*, *92*(12), e00515-18. https://doi.org/10.1128/JVI.00515-18

Kuraku, S., & Kuratani, S. (2006). Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. *Zoological Science*, *23*(12), 1053-1064. https://doi.org/10.2108/zsj.23.1053

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2020). CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Research*, *48*(D1), D265-D268. https://doi.org/10.1093/nar/gkz991

Malik, H. S., & Eickbush, T. H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Research*, *11*(7), 1187-1197. https://doi.org/10.1101/gr.185101

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, *24*(16), 1757-1764. https://doi.org/10.1093/bioinformatics/btn322

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

Pereira, A. M., Levy, A., Vukić, J., Šanda, R., Levin, B. A., Freyhof, J., Geiger, M., Choleva, L., Francisco, S. M., & Robalo, J. I. (2021). Putting european lampreys into perspective: A global-scale multilocus phylogeny with a proposal for a generic structure of the Petromyzontidae. *Journal of Zoological Systematics and Evolutionary Research*, *59*(8), 1982–1993. https://doi.org/10.1111/jzs.12522

Rodríguez-Varela, Gonzalo. (2021). *Relacións filoxenéticas dos retrovirus endóxenos de* Petromyzon marinus [Trabajo de fin de grado, Universidade da Coruña]. RUC. http://hdl.handle.net/2183/29253

RStudio. (2021). *RStudio Desktop* (Versión 1.4.1717) [Programa de ordenador]. http://www.rstudio.com

Smith, J. J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M. S., Yandell, M. D., Manousaki, T., Meyer, A., Bloom, O. E., Morgan, J. R., Buxbaum, J. D., Sachidanandam, R., Sims, C., Garruss, A. S., Cook, M., Krumlauf, R., Wiedemann, L. M., Sower, S. A., … Li, W. (2013). Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genetics*, *45*(4), 415-421. https://doi.org/10.1038/ng.2568

Smyshlyaev, G., Voigt, F., Blinov, A., Barabas, O., & Novikova, O. (2013). Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. *Proceedings of the National Academy of Sciences*, *110*(50), 20140-20145. https://doi.org/10.1073/pnas.1310958110

Stoye, J. P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews Microbiology*, *10*(6), 395-406. https://doi.org/10.1038/nrmicro2783

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022-3027. https://doi.org/10.1093/molbev/msab120

Telesnitsky, A. (2010). Retroviruses: Molecular biology, genomics and pathogenesis. *Future Virology*, *5*(5), 539-543. https://doi.org/10.2217/fvl.10.43

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Ustyantsev, K., Novikova, O., Blinov, A., & Smyshlyaev, G. (2015). Convergent evolution of ribonuclease H in LTR retrotransposons and retroviruses. *Molecular Biology and Evolution*, *32*(5), 1197-1207. https://doi.org/10.1093/molbev/msv008

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Dempsey, D. M., Dutilh, B. E., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Nibert, M., Rubino, L., Sabanadzovic, S., Simmonds, P., Varsani, A., … Davison, A. J. (2019). Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). *Archives of Virology*, *164*(9), 2417-2429. https://doi.org/10.1007/s00705-019-04306-w

Wang, J., & Han, G.-Z. (2021). A sister lineage of sampled retroviruses corroborates the complex evolution of retroviruses. *Molecular Biology and Evolution*, *38*(3), 1031-1039. https://doi.org/10.1093/molbev/msaa272

Wang, J., & Han, G.-Z. (2022). A missing link between retrotransposons and retroviruses. *mBio*, *13*(2), e00187-22. https://doi.org/10.1128/mbio.00187-22

Wei, X., Chen, Y., Duan, G., Holmes, E. C., & Cui, J. (2019). A reptilian endogenous foamy virus sheds light on the early evolution of retroviruses. *Virus Evolution*, *5*(1), vez001. https://doi.org/10.1093/ve/vez001

Xiong, Y., & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal*, *9*(10), 3353-3362. https://doi.org/10.1002/j.1460-2075.1990.tb07536.x

Xu, X., Zhao, H., Gong, Z., & Han, G.-Z. (2018). Endogenous retroviruses of non-avian/mammalian vertebrates illuminate diversity and deep history of retroviruses. *PLOS Pathogens*, *14*(6), e1007072. https://doi.org/10.1371/journal.ppat.1007072

Zheng, J., Wang, J., Gong, Z., & Han, G.-Z. (2021). Molecular fossils illuminate the evolution of retroviruses following a macroevolutionary transition from land to water. *PLOS Pathogens*, *17*(7), e1009730. https://doi.org/10.1371/journal.ppat.1009730

Zheng, J., Wei, Y., & Han, G.-Z. (2022). The diversity and evolution of retroviruses: Perspectives from viral “fossils”. *Virologica Sinica, 37*(1), 11-18. https://doi.org/10.1016/j.virs.2022.01.019

8. Anexo A

Anexo A: Tratamiento del Excel "Resultados BLAST.xlsx".

En primer lugar, se nombraron todas las columnas en el orden en el que se mostraban con los siguientes nombres: "Query acc. Ver", "Subject acc. Ver", "% identity", "Alignment length", "Mismatches", "Gap opens", "Q start", "Q end", "S start", "S end", "Evalue" y "Bit score". Posteriormente, se le añadieron columnas nuevas: "Cromosoma" (entre "Subject acc. Ver" y "% identity"), "Strand" (entre "S end" y "Evalue"), "Posición con respecto a rt", "el fragmento de la query a comprobar", "% de similitud de nt", "Inserciones fragmentadas" y "Notas" (estas últimas todas tras "Bit score").

Para empezar, se modificó la columna de "Cromosoma". Para ello, se anotó el número de cromosoma para cada inserción, comprobando que las coordenadas fueran las adecuadas. En el caso de que no hubiese el nivel cromosoma, se sustituyó el nombre de la columna por "Scaffold" y se anotó el número del scaffold.

A continuación, se completó la columna "Strand". En esta, se buscó detallar la orientación de la hebra que alineaba con la query, ya que el BLAST descarga sólo las hebras "+". Con este fin, se revisaron los alineamientos en el resultado proporcionado por la búsqueda del BLAST. En ellos, se buscó qué hebra estaba siendo utilizada para el alineamiento y se anotó en dicha columna. Aquellos casos en los que la hebra utilizada para el alineamiento era la "+", se le adjudicaba el número 1 en la tabla. En cambio, si la hebra utilizada era la "- ", se le adjudicaba el número 2 en la tabla. Para facilitar la visualización de estas observaciones se empleó una regla de color adjudicando el color verde a los números 1 y el color amarillo a los números 2.

Teniendo todo esto en cuenta, se revisaron los archivos FASTA con las inserciones en el programa BioEdit (Hall, 1999). Así, en aquellos cromosomas o scaffolds marcados de color amarillo fue necesario obtener la reserva complementaria de dicha cadena. Para ello, se emplearon las opciones sequence>nucleid acid>reverse complement.

El siguiente paso consistió en cubrir las columnas "el fragmento de la query a comprobar" y "% de similitud de nt". La primera columna sólo se cubrió si la longitud máxima del fragmento a comprobar no superaba los 10 nucleótidos. El encabezado de dicha columna se completó con los nucleótidos de dicho fragmento máximo. A continuación, para cada inserción se incluyó el fragmento específico de la query a comparar.

En ocasiones, el programa BLAST descarta algunos de los primeros nucleótidos para el alineamiento porque los considera muy diferentes. Con todo, a veces comete errores y se debe comprobar. Para ello, en primer lugar, se identificó qué alineamientos no empezaban en el primer nucleótido de la query. Para este fin, se ejecutó un formato condicional sobre la columna "Q start" destacando con un color azul las filas mayores que 1.

En los casos destacados en azul, se comprobó si los nucleótidos no utilizados para el alineamiento eran similares a los de la query. Si sólo dos nucleótidos (en el caso de que sean más de 3) eran diferentes a la query, se conservaban. En caso contrario, se descartaban destacándolos en color rojo para facilitar su visualización. Si se decidió conservar los nucleótidos, se procedió a modificar el hit con ayuda de BioEdit (Hall, 1999) y añadir dichos nucleótidos.

Los dos siguientes pasos se realizaron soló con las secuencias pertenecientes al dominio RH. En primer lugar, se cubrió la columna "Posición con respecto a RT". Para

ello, se comprobó la posición de las coordenadas de los hits del dominio RH con respecto al dominio RT para verificar si se encontraban antes o después de este (eliminando alguna que estuviera antes si no tenía sentido con respecto a la distribución de estos dominios en el genoma e incluyéndolas en inserciones fragmentadas).

A continuación, se cubrió la columna “Inserciones fragmentadas”. En esta se diferenciaron dos categorías:

- Las inserciones presentes en el dominio RT, pero ausentes en el dominio RH.
- Las inserciones presentes en el dominio RH, pero ausentes en el domino RT.

En todos los casos, se revisó su existencia y se anotaron los cromosomas o los cromosomas y las coordenadas. Esto último sólo en el caso de que, en algún cromosoma, estuvieran presentes unas coordenadas y otras no.

Por último, se cubrió la columna correspondiente a “Notas”. En ella, se incluyó el número de cromosomas diferentes del total de cromosomas de la especie descargados (sin tener en cuenta los cromosomas “duplicados” ya mencionados con anterioridad y el hecho de que haya más de una inserción por cromosoma).

Tabla A1

Caracterización de las inserciones pertenecientes al dominio RT presentes en el genoma de Petromyzon marinus.

Esta tabla está disponible en la hoja “RT_Petromyzon-marinus” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

Tabla A2

Caracterización de las inserciones pertenecientes al dominio RT presentes en el genoma de Lethenteron reissneri.

Esta tabla está disponible en la hoja “RT_Lethenteron-reissneri” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

Tabla A3

Caracterización de las inserciones pertenecientes al dominio RT presentes en el genoma de Lethenteron camtschaticum.

Esta tabla está disponible en la hoja “RT_Lethenteron-camtschaticum” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

Tabla A4

Caracterización de las inserciones pertenecientes al dominio RT presentes en el genoma de Entosphenus tridentatus.

Esta tabla está disponible en la hoja “RT_Entosphenus-tridentatus” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

Tabla A5

Caracterización de las inserciones pertenecientes al dominio RH presentes en el genoma de Petromyzon marinus.

Esta tabla está disponible en la hoja “RH_Petromyzon-marinus” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

Tabla A6

Caracterización de las inserciones pertenecientes al dominio RH presentes en el genoma de Lethenteron reissneri.

Esta tabla está disponible en la hoja “RH_Lethenteron-reissneri” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

Tabla A7

Caracterización de las inserciones pertenecientes al dominio RH presentes en el genoma de Lethenteron camtschaticum.

Esta tabla está disponible en la hoja “RH_Lethenteron-camtschaticum” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

Tabla A8

Caracterización de las inserciones pertenecientes al dominio RH presentes en el genoma de Entosphenus tridentatus.

Esta tabla está disponible en la hoja “RH_Entosphenus-tridentatus” del Excel presente en el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EZHahW4p3UBEuCKcJ-Xf2rABK9nlCA7xqLCF_WL-PGky2A?e=rFymoQ

9. Anexo B

Anexo B: Modificación de los nombres de las secuencias con la finalidad de conseguir un nombre corto e identificable.

Para modificar los nombres de las secuencias se utilizaron los programas RStudio (RStudio, 2021) y Excel a lo largo del proceso detallado a continuación:

1. Se guardó el archivo “erv1_dominio_ all.fas” en formato XML.

2. Con ayuda de un convertidor, se transformó el archivo en formato XML en un archivo en formato CSV, siempre bajo el mismo nombre.

3. Se abrió el programa RStudio (RStudio, 2021), estableciendo como directorio la carpeta donde se estaban guardando todos los archivos.

4. Se abrió el archivo CSV y se realizaron las primeras modificaciones con ayuda de la herramienta de búsqueda y reemplazo:

   a. Se eliminó la fila “description,phylogeny,sequence”.

   b. Se reemplazaron los “:” que separaban el “accession number” y las coordenadas por un espacio.

   c. Se sustituyó la “,” que seguía al número del cromosoma o scaffold por “_”. En los casos en los que no estaba seguido por “,” se sustituyó la frase que había detrás de él por “_”.

   d. Se cambiaron las “,,” tras el nombre de la secuencia por un espacio.

5. Una vez realizados todos estos cambios se guardó de nuevo el archivo.

6. Se separaron los nombres de las secuencias y las propias secuencias en columnas buscando que el “accession number”, las coordenadas, el género, el epíteto específico de la especie, el cromosoma, la secuencia y el resto de información separada por espacio quedase en distintas columnas. Para ello, se utilizó el siguiente comando modificándolo según fue necesario (cambiando el nombre del dominio y la especie):

```
>RT_Pet <- read.delim("Hits-RT-PetMar.csv", header=FALSE, sep=" ")
```

7. Se duplicó la columna de las coordenadas con la finalidad de colocarla detrás del número del cromosoma. Para ello, se empleó el siguiente comando, sustituyendo V12 por la columna no relevante que se haya situada detrás del número del cromosoma:

```
> RT_Pet$V12 = RT_Pet$V2
```

8. Se modificó la especie para ajustarla a “tres primeras letras del género + tres primeras letras del epíteto específico” con la ayuda de los siguientes comandos, cambiándolos según la especie:

```
> RT_Pet$V3[grepl("Petromyzon", RT_Pet$V3)] <- "pet"

> RT_Pet$V4[grepl("marinus", RT_Pet$V4)] <- "mar_"
```

9. Finalmente, se eliminaron las columnas que sobraban con ayuda del comando:

> RT_Pet$V2 = NULL (cambiando V2 según la columna a eliminar).

10. Se guardó el archivo en formato txt con ayuda del comando:

> write.table(RT_Pet, "Hits-RT-Petmar-nombres.txt", sep=";")

11. Se abrió Excel y se cargó el archivo en formato txt (Datos>Obtener datos externos desde un archivo de texto).

12. Se eliminaron las columnas que sobraban.

13. Como se quería que las columnas que formaban el nombre estuviesen en una sola, se añadió una columna detrás de la de las coordenadas y en ella, con ayuda de la función CONCATENAR, se unieron las columnas correspondientes al nombre (letras de la especie, número de cromosoma y coordenadas).

14. Se copiaron la columna de concatenación y la de las secuencias y se llevaron a un convertidor de un archivo txt en un archivo formato FASTA (https://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html). Este archivo FASTA se guardó de nuevo bajo el nombre “Hits-dominio-tres primeras letras de género y epíteto específico.fas”.

Tabla B1

Descripción de las secuencias facilitando el nombre que presentarán en los archivos de alineamiento de secuencias y árboles filogenéticos.

Esta tabla está disponible en el Excel al que remite el siguiente enlace: https://udcgal.sharepoint.com/:x:/s/TFGERVsPetromyzon/EeRw_VS99flBrVcMEfu_mK4BHy5Am4qSZWuyoYl2PnVt_w?e=gUKvds

10. Anexo C

Anexo C: Revisión del alineamiento realizado por el programa Clustal W (Thompson et al., 1994) contenido en BioEdit (Hall, 1999).

A lo largo del alineamiento, el programa puede cometer “errores” (alinear las secuencias de una forma considerada errónea por nosotros o ignorar ciertos aspectos que debería tener en cuenta).

Entre ellos, se buscó la presencia de posiciones CpG. En estas, la posición C está metilada y es propensa a la desaminación espontánea. De manera que, se puede pasar de tener CG a tener TG o AC. Este tipo de posiciones generan “ruido” ya que provocan que secuencias que han divergido hace mucho tiempo coincidan en ciertas posiciones haciéndolas parecer similares. Por lo tanto, si se observa un exceso de mutación, se deben eliminar.

Por otro lado, también, se buscaron posiciones en las que tuviera cabida otra forma de alineamiento no considerada por el programa. En esos casos, se añadieron o quitaron gaps en lugares determinados para asegurar un mejor alineamiento.

Dataset C1

Alineamiento de las inserciones del dominio RT presentes en Petromyzon marinus, Lethenteron reissneri, Lethenteron camtschaticum y Entosphenus tridentatus.

Este alineamiento está disponible en el siguiente enlace: https://udcgal.sharepoint.com/:u:/s/TFGERVsPetromyzon/ETL5Oj5oXYxGklY3GTabaZQBVQhlnYuc84n6bwKBbRYa4A?e=D7SPeK

Dataset C2

Alineamiento de las inserciones del dominio RH presentes en P. marinus, L. reissneri, L. camtschaticum y E. tridentatus.

Este alineamiento está disponible en el siguiente enlace: https://udcgal.sharepoint.com/:u:/s/TFGERVsPetromyzon/ETus1V3hEdVFsAps0bdsErUBWSySO7ALbwd1pAG4tMMokA?e=x1LdWq

Dataset C3

Alineamiento de las inserciones del dominio Tether dentro del dominio RH presentes en P. marinus, L. reissneri, L. camtschaticum y E. tridentatus.

Este alineamiento está disponible en el siguiente enlace: https://udcgal.sharepoint.com/:u:/s/TFGERVsPetromyzon/EZPGMKSAnUxBlQHPsRa0jd0BMpX9C3YzVUt6wPKr3oL9wQ?e=8yqBex

Dataset C4

Alineamiento de las inserciones del nuevo dominio RH dentro del dominio RH presentes en P. marinus, L. reissneri, L. camtschaticum y E. tridentatus.

Este alineamiento está disponible en el siguiente enlace: https://udcgal.sharepoint.com/:u:/s/TFGERVsPetromyzon/EaRXW-FwCWlNmy3pj804CSwBg8vpkWsV3BpZpxnmxLZxfQ?e=bV79A2

11. Anexo D

Anexo D: Descripción detallada de la elaboración y modificación de la secuencia consenso

Antes de elaborar la secuencia consenso, se tuvo en cuenta que el programa MEGA 11(Tamura et al., 2021), el cual será utilizado para construir las filogenias, no detecta ningún carácter que no sea A/T/G/C/_.

¿Por qué se prestó atención a esto?

Porque, según el umbral de frecuencia empleado, el programa BioEdit (Hall, 1999) puede tener problemas para decidir qué estado (A/T/G/C) es más frecuente. En esos casos, utiliza una letra que designe todos los nucleótidos presentes en esa posición siguiendo los criterios de la International Union of Pure and Applied Chemestry (IUPAC).

En consecuencia, se tomaron las siguientes decisiones:

Se revisó el umbral de frecuencia (options>preferences>consensus) y se puso al 55%.

Una vez obtenida la secuencia consenso, se revisó para modificar aquellas posiciones en las que aparecía una letra diferente a las 4 habituales sustituyéndola por el nucleótido más frecuente. Para saber cuál era el nucleótido más frecuente, se creó una segunda secuencia consenso bajando el umbral de frecuencia al 50%.

Además de tener problemas para decidir qué nucleótido es más frecuente, BioEdit (Hall, 1999) también puede tener dificultades para colocar un gap (“_”) en la secuencia consenso si este es el estado más frecuente en la posición estudiada. Por ello, fue necesario revisar la secuencia consenso y sustituir A/T/C/G por un gap donde fue preciso. Es muy importante que estos estén presentes ya que su ausencia genera proteínas truncadas.

Las secuencias consenso de cada alineamiento están presentes en los datasets del Anexo C.

12. Anexo E

Anexo E: Separación del dominio RH en el dominio Tether y el nuevo dominio RH.

En primer lugar, se realizó un BLAST sobre la secuencia consenso (eliminando los gaps resultado del alineamiento) utilizando como query el dominio Tether. Esta query fue creada de la misma forma que en el apartado 3.1 de “Material y Métodos”. De esta forma, a través del alineamiento, se obtuvieron las coordenadas del dominio Tether en la secuencia consenso.

A continuación, se abrió solo la secuencia consenso en el BioEdit (Hall, 1999) y se localizaron esas coordenadas. Seguidamente, se anotaron los 10 primeros nucleótidos de estas. Estos fueron buscados en la secuencia consenso dentro del archivo del alineamiento.

Una vez localizado el dominio Tether en el alineamiento, se extrajeron las porciones de las secuencias correspondientes a las coordenadas de dicho dominio. De esta manera, se obtuvo un archivo formado únicamente por las secuencias correspondientes al dominio Tether.

Tras todo esto, se repitió el mismo procedimiento para el nuevo dominio RH, simplemente extrayendo la parte de las secuencias a continuación de las coordenadas del dominio Tether. En ambos casos, se guardó el archivo bajo el nombre “erv1_dominio_all_aln.fas”, dichos archivos se encuentran en los Datasets C3 y C4 del Anexo C.

Al extraer unas determinadas posiciones de unas secuencias, al crear el nuevo archivo, el programa les modifica el nombre agregándole la posición. En consecuencia, de nuevo habría que modificarles la nomenclatura.

Con esa finalidad, se repitieron los apartados 1-3 del Anexo B. Posteriormente, simplemente, se eliminaron las posiciones con ayuda de la herramienta buscar y reemplazar. A continuación, se guardó el nuevo CSV y se repitió el punto 14 de dicho Anexo B.

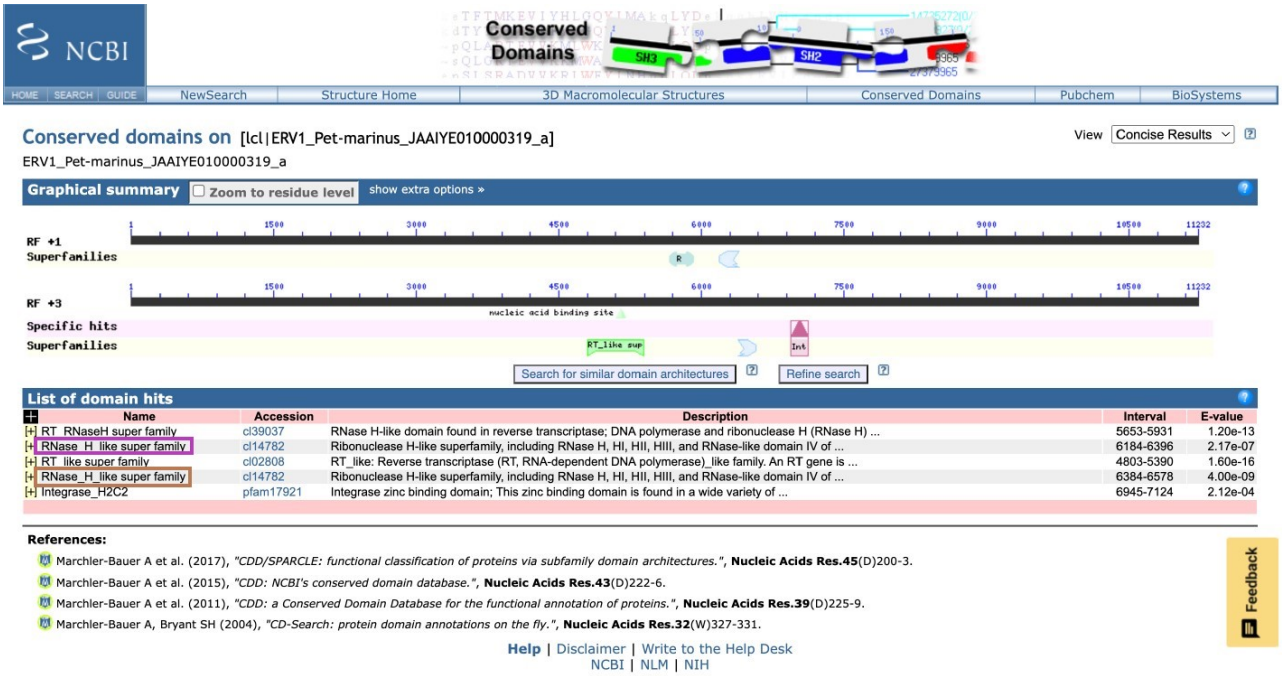


Figura E1. Caracterización de los dominios presentes en ERV1 JAAIYE01000319_a obtenido a partir del TFG de Rodríguez-Varela (2021) y utilizado como query. Se observan los dominios estudiados RT, Tether y nuevo dominio RH.

13. Anexo F

Anexo F: Árbol filogenético que evidencia la adecuación de las dos inserciones petmar_68 como outgroup

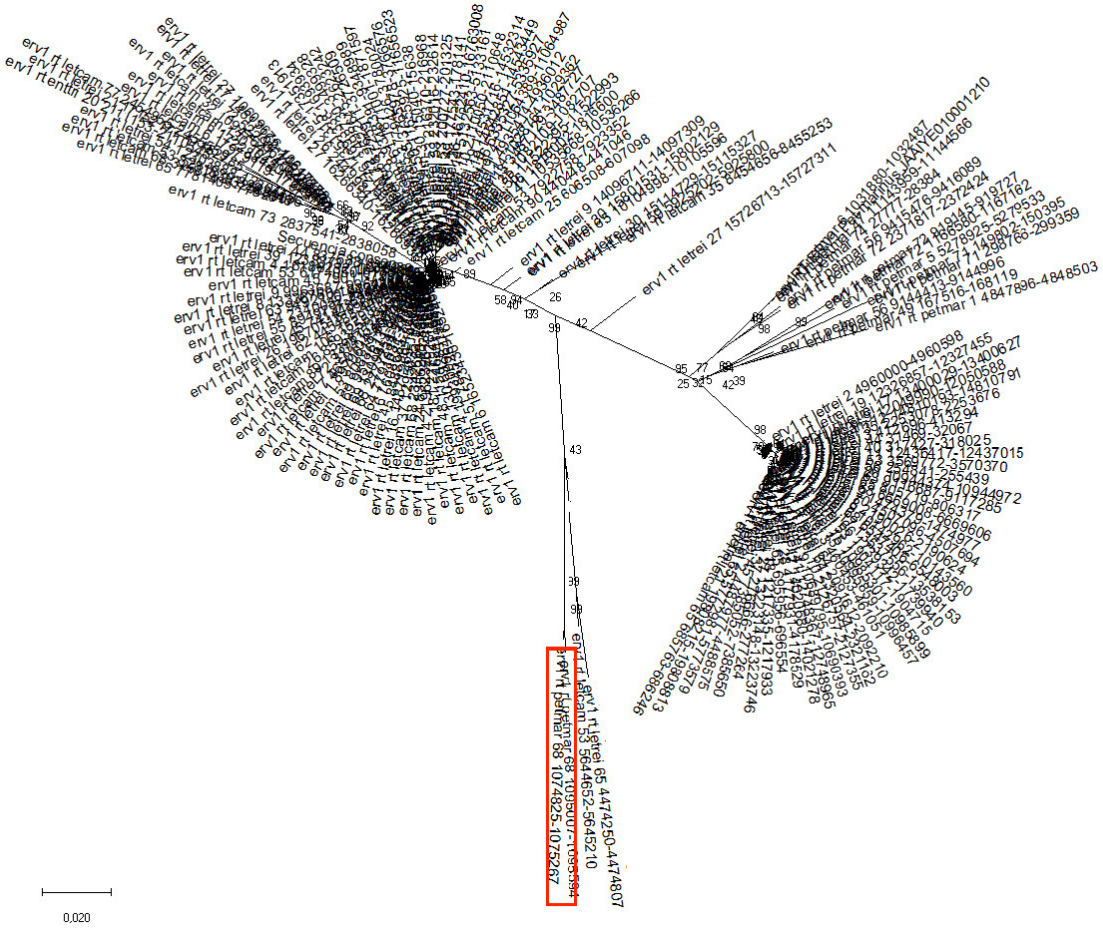


Figura F1. Filogenia con disposición radial elaborada en base al conjunto de secuencias "erv1_rt_all_aln.fas" utilizando como método Neighbor-Joining bajo las condiciones de la Figura 5B. Se destaca en rojo el par de secuencias petmar_68 consideradas como las más aptas para ser empleadas como outgroup.