

UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

**“Los retropseudogenes de argininosuccinato sintetasa en el genoma humano.
Análisis genómico y evolutivo”**

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G A

P R E S E N T A

BIBIANA RODRÍGUEZ PONCE

DIRECTOR DE TESIS: DR. VÍCTOR MANUEL VALDÉS LÓPEZ



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Hoja de datos de jurado

<p>1. Datos del alumno Rodríguez Ponce Bibiana 53 04 10 94 Universidad Nacional Autónoma de México Facultad de Ciencias Biología 09855382-6</p>
<p>2. Datos del Tutor Dr. Víctor Manuel Valdés López</p>
<p>3. Datos del sinodal 1 Dra. María Alicia González Manjarrez</p>
<p>4. Datos del sinodal 2 Dr. Luis José Delaye Arredondo</p>
<p>5. Datos del sinodal 3 Dr. Arturo Carlos II Becerra Bracho</p>
<p>6. Datos del sinodal 4 M. en C. Alfonso José Vichis Peluyera</p>

7. Datos del trabajo escrito

Los retropseudogenes de argininosuccinato sintetasa en el genoma humano.

Análisis genómico y evolutivo.

83 p

2006

A mis padres

Por darme una tierra fértil de amor y apoyo donde crecer

A mi hermana, mi sobrino y cuñado

Por darle luz a nuestra pequeña familia y por todas sus porras

A todos mis amigos

Por ser mi segunda familia y estar aquí, con migo, aunque sea en
pensamiento
Por su ayuda, las pláticas espesas, las risas, lagrimas y claro esta, por todas
las chelas compartidas

Agradezco

Al Dr. Víctor Valdés López

Por todo el conocimiento, apoyo y comprensión que ha tenido a bien
brindarme

Por darme la oportunidad de pertenecer a un gran equipo de trabajo y en
especial por su cuasi infinita paciencia a todos mis ¡¡¡¡Oye Víctor!!!!

A mis compañeros y amigos del Laboratorio de Biología Molecular
Alfonso Vilchis, Claudia Segal y Beatriz Rodarte

Por todas las cosas académicas y personales que han compartido conmigo

A mis sinodales por su entusiasmo y todos sus acertados comentarios

A los maravillosos profesores que me han mostrado la belleza del
pensamiento

Índice

Resumen	1
Introducción	2
Elementos transponibles.....	5
Pseudogenes.....	7
Globinas.....	12
Argininosuccinato sintetasa.....	14
I. Justificación	17
II. Antecedentes	19
III. Objetivos	20
a. General.....	20
b. Específicos.....	20
IV. Metodología	21
a. Obtención y caracterización de las retrosecuencias.....	21
b. Datación de las secuencias.....	24
c. Potencial codificantes.....	25
d. Análisis de polimorfismos en las secuencias.....	27
V. Resultados y discusión	28
a. Búsqueda de retropseudogenes en la superfamilia de las hemoglobinas.....	28
b. Descripción general y búsqueda de las retrosecuencias para ASS.....	30
c. Descripción del cDNA.....	32
d. Búsqueda de eventos de duplicación.....	37
e. Datación de los retropseudogenes.....	39
f. Determinación del potencial codificante de las retrosecuencias.....	42
g. Análisis de polimorfismos en las secuencias.....	44
h. Análisis de la vecindad metabólica de la ASS.....	47
i. Análisis de la vecindad evolutiva de la ASS humana.....	59
Conclusiones y perspectivas	66
Anexos	70
Bibliografía	76

Resumen

Los retropseudogenes (retros) son secuencias no funcionales generadas por procesos de reversotranscripción y reintegradas al genoma. Este tipo de secuencias pueden servir para darnos una panorámica de los procesos pasados ocurridos dentro del genoma, cómo cambios en los patrones de expresión de genes. En el presente trabajo se analizaron los retropseudogenes asociados con la enzima, participante del ciclo de la urea y óxido nítrico, argininosuccinato sintetasa (ASS) en el genoma humano, encontrando para ésta, un número de 14 retrosecuencias –con una antigüedad estimada de entre 8.9 y 17.8 millones de años (ma)– ninguna de las cuales parecen estar sujetas a presiones de selección. Posteriormente, con ayuda de los retropseudogenes se reconstruyó la variante del mensajero que les dio lugar, así como también se localizó dentro de las secuencias actuales de humano y chimpancé los cambios linaje específicos que sufrieron posteriores a la generación de los retros. Al analizar la vecindad metabólica de la ASS se encontró que solo el gen de esta enzima generó retrosecuencias, debido posiblemente a su papel como enzima nodo (tasa limitante). Por último, el análisis de su vecindad evolutiva mostró que humano y chimpancé comparten, como se había esperado, todos los retropseudogenes de ASS, mientras que el ratón muestra un patrón de generación de retrosecuencias muy diferente, con solo una retrocopia que dadas sus características podría ser funcional.

Introducción

Hace 30 años Frederick Sanger desarrolló una metodología para poder secuenciar material genético (la técnica de dideoxi), después de que en 1953 hubiese obtenido la primera secuencia completa de una proteína, la insulina. Al ir perfeccionando sus metodologías pudo en 1977 secuenciar el genoma, es decir la colección completa de material genético presente en un organismo particular, del bacteriofago ϕ x 174, el cual se convirtió en el primer organismo secuenciado completamente, con solo 5,375 nucleótidos (nt). A pesar de que en la actualidad esta cifra parezca mínima, para ese momento fue uno de mayores logros en el desarrollo de la biología molecular y la base de lo que en la actualidad se llama la era genómica.

Posteriormente el mismo Sanger secuenció el genoma completo del DNA presente en la mitocondria humana y el del bacteriofago λ de 16,338 y 48,500 nt respectivamente. A partir de este momento los científicos se dieron cuenta del alcance de tener un genoma completo y sus repercusiones en el estudio de los organismos. Así utilizando como base las técnicas de Sanger se comenzaron a secuenciar genomas de virus, dado su pequeño tamaño y posteriormente de bacterias con importancia médica. Para 1990 ya se habían secuenciado una gran cantidad de genomas virales y se tenía planeado comenzar a secuenciar bacterias y otros organismos con genomas de mayor tamaño. El desarrollo de nuevas metodologías así como los avances computacionales permitieron pensar en la secuenciación del genoma del ser humano, naciendo así el proyecto genoma humano, el cual proyectaba tener listo el mapa completo del humano para el año 2005.

Durante los siguientes 10 años se continuaron desarrollando metodologías que hicieron posible tener el borrador completo del genoma humano para febrero del 2001, publicado por los dos grupos que trabajaron sobre ello, el Consorcio Internacional de Secuenciación del Genoma Humano (IHGSC) –quienes publicaron en la revista Nature– y el grupo encabezado por Craig Venter –los cuales lo hicieron en la revista Science– (IHGSC, 2001 y Venter *et al*, 2001).

Al contar con una gran cantidad de información acerca de genes y genomas, se tuvieron que desarrollar bases de datos (tanto públicas como privadas) que albergaran dichas secuencias y toda aquella información relacionada con ellas así como recursos computacionales que permitieran su estudio. De esta manera surgieron sitios web muy importantes como el Centro Nacional para la Información Biotecnológica (NCBI, www.ncbi.nlm.nih.gov) creado en 1988, el ENSEMBL (www.ensembl.org) fundado por el Wellcome Trust Sanger Institute y el Instituto Europeo de Bioinformática, la Enciclopedia de Kyoto de Genes y Genomas (KEGG, <http://www.genome.jp/kegg/>), entre otros. Hasta la fecha el NCBI reporta 1665 genomas secuenciados solamente de virus, 354 genomas de arqueobacterias y eubacterias y 44 de eucariontes. Entre los genomas de vertebrados completamente secuenciados se tiene el genoma de humano, rata, ratón, chimpancé, perro, gallo y toro.

Según el Consorcio Internacional del Genoma Humano, dicho genoma está conformado por alrededor de 3.08×10^9 pares de bases (pb). Éste consorcio identificó 19,599 genes conocidos pero se cree que puede haber entre 20,000 a 25,000 genes que codifiquen proteínas. En el linaje del humano, al igual que en otros grupos, se han dado casos de nacimiento de nuevos genes, existiendo 1,183 genes que presentan divergencia con el gen original, dando así lugar a la formación de familias multigénicas. También se ha dado el caso de pérdida o “muerte” de genes (IHGSC, 2004).

Ejemplo de pérdida de genes –por conversión a pseudogenes o eliminación– se da en la familia de los receptores olfativos humanos, la cual en la actualidad cuenta con 388 genes funcionales y 414 pseudogenes, pero se cree que ha perdido aproximadamente 430 genes desde que compartiera el último ancestro con el ratón (el cual se estimó que tenía aproximadamente 750 genes para receptores olfativos). De manera contraria, el ratón ganó aproximadamente 350 genes olfativos, contando en la actualidad con 1037 genes funcionales y 354 pseudogenes (Niimura y Nei, 2005).

Al comparar el genoma humano con el del chimpancé, los cuales compartieron un último ancestro común entre 6 y 10 millones de años (ma), se observó que la diferencia en sus secuencias de amino ácidos es menor de 1% (Varki, 2000). El análisis indica que el 95% del genoma del chimpancé puede ser alineado con segmentos correspondientes en el genoma humano, siendo la divergencia en estas regiones de

aproximadamente 1.2%. Además el 5% no alineado corresponde a deleciones e inserciones ocurridas en ambos genomas desde la separación de sus linajes (Olson y Varki, 2003).

Por su parte, el genoma del ratón, es más pequeño que el genoma humano (2.5×10^9 contra 3.08×10^9 pb); a pesar de esto, las predicciones realizadas hasta el 2002, mostraban que ambos genomas contenían un número similar de genes que codifican proteínas (aproximadamente 30,000). El análisis de este genoma muestra que la proporción de genes específicos de roedores es menor al 1%, mientras su proporción de genes específicos de mamífero es de aproximadamente 14%. Cerca del 90% del genoma del ratón presenta sintenia (es decir, conserva la misma posición de los genes en el genoma) con el genoma humano, teniendo segmentos en los cuales se ha conservado el orden de los genes desde el último ancestro común de las dos especies. Además el 80% de las proteínas del ratón presenta un ortólogo estricto (genes que pasaron a los dos linajes desde su divergencia con el ancestro común) en una relación 1:1 con el humano, sin embargo dichas proteínas son codificadas por familias multigénicas con diferentes proporciones de genes en ambos genomas, perdiéndose con ello la relación 1:1 (MGSC, 2002).

En el genoma del ratón docenas de familias génicas locales han sufrido expansión, en especial las relacionadas con la reproducción, inmunidad (como es el caso de la familia de las citocromo P450) y olfato. Esto sugiere que estos sistemas fisiológicos han sido el centro de innovaciones linaje-específicas en los roedores. Otro punto de interés es la variabilidad en la tasa de cambio en el genoma del ratón, ya que algunos grupos de genes evolucionaron a tasas más aceleradas que otros. Tampoco se observa uniformidad en las tasas de cambio y transposición al compararse contra el genoma humano (MGSC, 2002).

En la actualidad se sabe que en el genoma humano alrededor del 2% es codificante, 33% esta conformado por secuencias no repetidas, entre las cuales se encuentran los pseudogenes, mientras más del 45% esta conformado por secuencias altamente repetidas.

Estas últimas pueden ser agrupadas en cinco tipos:

- 1) Repeticiones derivadas de transposones (secuencias interdispersas).
- 2) Retrocopias inactivas de genes, denominadas retropseudogenes o pseudogenes procesados.
- 3) Repeticiones de secuencias simples como $(A)_n$, $(CA)_n$ o $(CGG)_n$.
- 4) Duplicaciones de segmentos del genoma de 10 a 300kb.
- 5) Secuencias repetidas en tandem, presentes en centrómeros, telómeros, brazos cortos de cromosomas acrocéntricos y grupos de genes ribosomales (IHGSC, 2001).

Elementos transponibles

En mamíferos, los elementos transponibles pueden ser agrupados en cuatro categorías, secuencias LINE (Long INterspersed Elements), SINE (Short INterspersed Elements), retro-transposones LTR y transposones de DNA. Los tres primeros se transponen a través de un intermediario de RNA mientras el cuarto lo hace directamente (IHGSC, 2001).

Las secuencias LTR (long terminal repeat) son estructuras que guardan un gran parecido con retrovirus. Estos elementos cuentan con secuencias regulatorias en ambos extremos. Al igual que los retrovirus contienen los genes *gag* y *pol* que codifican para la capsida viral y una transcriptasa reversa, ribonucleasa e integrasa, pero a diferencia de las partículas virales las secuencias LTR han perdido o tienen dañado el gen *env* cuya proteína facilita el movimiento del virus de una célula a otra, quedando restringido a sólo moverse dentro del genoma de una célula (Deininger & Batzer, 2002).

Las secuencias LINE están ampliamente distribuidas en los genomas de mamíferos, en el humano comprenden aproximadamente el 17% del genoma, presentando aproximadamente 520,000 copias de las cuales el 99.8% no son funcionales. En su estado funcional presentan una longitud de 6Kb, son capaces de

retro-transponerse autónomamente a través de un intermediario de RNA. Codifican dos marcos de lectura abiertos (ORFs) separados por 63 pb, el primero genera una proteína de unión al RNA de 40kDa, mientras la segunda proteína presenta actividades de endonucleasa y reverso-transcriptasa con un peso de 150kDa. Los dos ORFs están precedidos por una región 5' no traducida, la cual contiene un promotor para la RNA polimerasa II y una región 3' no traducida que presenta una cola de poli(A⁺) (Han *et al*, 2005).

Según Han y colaboradores en 2005, las proteínas derivadas de LINE exhiben una marcada preferencia *cis*, es decir, usan preferentemente al mRNA de LINE como molde y el DNA resultante lo reinsertan al genoma, pero autores como Esnault y colaboradores en 2000, afirman que también pueden presentar actividad *trans*, lo cual implica que pueden tomar RNAs mensajeros no asociado con LINE, retrotranscribirlos y reintegrarlos al genoma, aunque esta actividad es menor a la *cis* debido posiblemente a un reconocimiento directo de los mRNA de LINE. Este hecho, como se verá más tarde, tiene amplias repercusiones en la dinámica del genoma.

Gracias a su amplia actividad las secuencias LINE están cambiando constantemente. Se sabe que a lo largo de la evolución de los primates las secuencias LINE han participado activamente en el reacomodo del genoma, teniendo un papel vital en la inactivación de genes así como en la delección de material genético, removiendo cerca de 7.5Mb desde la radiación del orden de los primates, dichas pérdidas pueden ser rastreadas en el genoma de humanos y chimpancés (Han *et al*, 2005). Se ha observado también que en organismos como *Drosophila melanogaster* la presencia de secuencias LINE dentro de sitios específicos del genoma sirven para estabilizar a los telómeros (Kazazian, 2004).

Por otro lado, las secuencias SINE pueden ser consideradas como secuencias no autónomas, dado que no cuentan con la maquinaria necesaria para retrotranscribirse. En los primates las secuencias SINE más abundantes son las llamadas secuencias Alu, las cuales, constituyen el 11% del total del genoma humano (Roy-Engel *et al*, 2002). Las SINE presentan generalmente entre 90-300 pb, son transcritas por la RNA polimerasa III, se cree que estas secuencias fueron derivadas de genes de tRNA o del 7SL RNA. Se cree que su movilización está mediada por la maquinaria de las secuencias LINE,

evidencia de ello es la alta identidad en las regiones 3'UTR (no traducida) de SINE y LINE, así como la presencia de una cola de poliA de tamaño parecido a la de las secuencias LINE (Deininger & Batzer, 2002).

Al igual que las secuencias LINE, las secuencias SINE están sujetas a constante cambio por mutaciones, entrecruzamiento, inserciones y deleciones, así también contribuyen al reacomodo y reestructuración del material genético y en algunos casos las han catalogado como secuencias mutagénicas, por inducir cambios en los patrones de expresión de genes en ensayos *in vivo* (Roy-Engel *et al*, 2002)

Al contar con elementos regulatorios propios, las secuencias LINE pueden alterar los patrones de expresión de genes, así también las SINE pueden llegar a tener la misma función pero al integrarse a regiones regulatorias y alterar la unión de estas con la RNA polimerasa o con sus proteínas asociadas. Otros mecanismos por los cuales ambos tipos de secuencias afectan la regulación de genes son: por la alteración de los sitios de poliadenilación e incorporándose a la región codificante del gen, alterando con ello el marco de lectura y la función del producto (Deininger & Batzer, 2002). Por ultimo se cree que este tipo de secuencias pueden estar implicadas en la inactivación del cromosoma X en mamíferos, aunque este mecanismo no está del todo comprendido, y que pudieron dar origen a la formación de los telómeros (Deininger & Batzer, 2002).

Pseudogenes

El genoma de los organismos eucariontes esta conformado por genes que se transcriben y/o traducen, pudiendo estar en copias simples, formar parte de una familia génica o presentar varias copias de si mismos. Además, en los genomas se tienen secuencias altamente repetidas y secuencias no codificantes como los llamados pseudogenes, entre otras.

Un pseudogen es una secuencia no funcional presente en el genoma de una población que presenta una alta similitud con uno o más genes homólogos (Mighell *et al*, 2000). La pérdida de dicha funcionalidad puede ser causada por la pérdida de promotores o elementos regulatorios y/o la acumulación de mutaciones que pueden alterar el marco de lectura (Zhang *et al*, 2003). Un rasgo importante de los pseudogenes

es que su secuencia de nucleótidos difiere de su contraparte funcional en puntos cruciales. No existe una notación específica para los pseudogenes, algunos autores anteponen al nombre del gen la letra griega ψ , en otros casos se coloca una P después del nombre del gen (Mighell *et al*, 2000).

Se han identificado diferentes tipos de pseudogenes en una gran variedad de organismos procariontes y eucariontes como: plantas insectos y vertebrados. Estos diferentes tipos pueden ser agrupados en tres categorías (D'Errico *et al*, 2004):

- Pseudogenes no procesados
- Pseudogenes mitocondriales
- Pseudogenes procesados o retropseudogenes

Pseudogenes no procesados

Los pseudogenes no procesados se generan por: 1) duplicación de un gen y una posterior inactivación, 2) duplicación de otro pseudogen preexistente como en el caso de $\psi\beta^x$ y $\psi\beta^z$ en la familia de las β -globinas ó 3) por la inactivación de un gen funcional (Li y Graur, 2000).

Los pseudogenes no procesados están por lo regular cercanos al gen que los originó. Además, pueden presentar regiones reguladoras e intrones como sus contrapartes funcionales (D'Errico *et al*, 2004). El proceso de duplicación se puede dar de las siguientes maneras: a) duplicación completa del gen, b) duplicación parcial del gen, c) duplicación completa del cromosoma, d) duplicación parcial del cromosoma y e) poliploidía o duplicación completa del genoma (Li y Graur, 2000).

Hardison en 1998, a través del estudio de la superfamilia de las globinas en diferentes grupos de organismos procariontes y eucariontes, demostró que la duplicación es uno de los procesos mediante los cuales los genomas ganan complejidad y se generan familias multigénicas a partir de un gen ancestral. Como se mencionó anteriormente, estas nuevas secuencias al no estar sujetas a presiones de selección

pueden tomar tres caminos: retención de la función original, adquisición de una nueva función ó pérdida de la función (Li y Graur, 2000).

Pseudogenes mitocondriales

Se cree que el empequeñecimiento del genoma mitocondrial es una consecuencia de la transferencia de genes de la mitocondria al genoma nuclear. Este proceso de integración de fragmentos de DNA mitocondrial (mtDNA) al DNA nuclear presumiblemente comenzó desde que los primeros endosimbiontes se establecieron como organelos. Esta teoría de desplazamiento de genes se ve respaldada por la presencia de genes funcionales, que codifican para proteínas de la mitocondria, dentro del genoma nuclear. Dichas transferencias se pueden explicar a través de dos mecanismos: transferencia directa de DNA, por recombinación no homóloga de fragmentos de mtDNA con DNA cromosomal ó por una transferencia mediada por RNA.

Pero no todas las integraciones dieron origen a genes funcionales, originándose los llamados pseudogenes mitocondriales. Este tipo de pseudogenes constituyen una alta proporción del genoma de mamíferos y pájaros. Se han encontrado algunos en *D melanogaster* y *C elegans*, aunque son casi inexistentes en peces (D'Errico *et al*, 2004). Woischnik y Moraes en 2002 encontraron 612 fragmentos de mtDNA en el genoma humano provenientes de integraciones independientes, mientras Zhang *et al*, en 2003 reportaron la presencia de 254 pseudogenes mitocondriales, recientemente adquiridos, en dicho genoma.

Pseudogenes procesados o retroseudogenes

Un retrogen o gen procesado es el resultado de un proceso de retrotransposición, es decir, de la transcripción reversa de un mRNA maduro seguido por la integración de este al DNA genómico (ver Figura 1); este proceso ocurre en cualquier célula, pero para poder fijarse en la población debe ocurrir en las células de la línea germinal. Las características principales de los retroseudogenes son: la total ausencia de intrones, la presencia, en algunos casos, de pequeñas regiones flanqueantes repetitivas y de una **cola de poli A** en el extremo 3' terminal (Zhang *et al*, 2003). Se puede nombrar a los retroseudogenes de diversas maneras como: *retros*, *copias retroprocesadas*, *pseudogenes procesados*. Si aun no se ha determinado su carácter de pseudogen se puede decir que se trata de una *retrosecuencia* o una *retrocopia*.

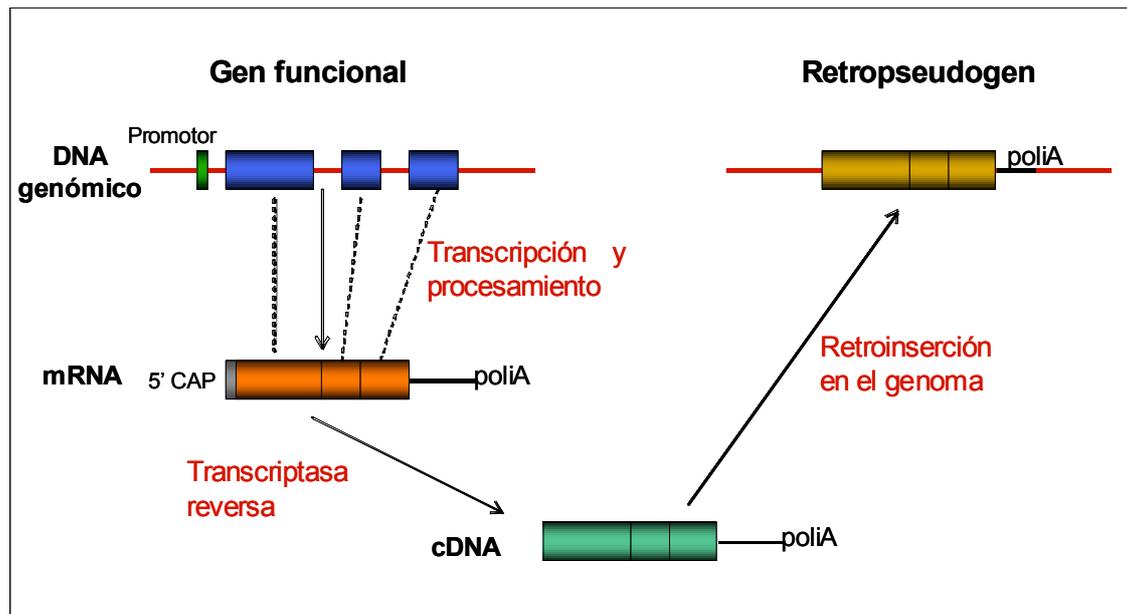


Figura 1. Mecanismo de generación de retrosecuencias (Tomada de Valdés-López *et al*, 2004)

Existen varias razones por las cuales las retrosecuencias pueden diferir de la secuencia original. En primer lugar, el proceso de retrotranscripción puede generar errores por lo cual el cDNA generado puede diferir del mRNA molde. En segundo lugar, el retrogen no presenta elementos reguladores, puesto que las regiones promotoras normalmente no son transcritas, a menos que el gen sea transcrito por la RNA polimerasa III, aunque esta solo transcribe a los RNA de transferencia. En tercer lugar, la retrosecuencia puede ser insertada en regiones sin promotor o elementos que le pudieran permitir transcribirse (Li y Graur, 2000). En general se sabe que esas copias no están sujetas a presiones de selección, siendo su tasa de mutación mucho mayor a su contraparte funcional y puede ir acumulando mutaciones o disrupciones que alteren el marco de lectura. Dadas estas razones la mayoría de las retrosecuencias no son funcionales nombrándolas retropseudogenes (Zhang *et al*, 2003).

Se ha calculado que sólo el 10% de los genes en el genoma humano presentan retrocopias ¿Pero qué características presentan estos genes que los hacen propensos a tener retrocopias? Los genes constitutivos son más propensos a generar retropseudogenes (Mighell *et al*, 2000). Entre estos tenemos a las proteínas de unión DNA/RNA, receptores de varios tipos, enzimas y proteínas ribosomales. Estas últimas son las que presentan el mayor número de retrocopias distribuidas en todo el genoma (Zhang *et al*, 2003), también son generados por genes que generan proteínas de

respuesta a estrés como la HSP10 Y HSP60 (Barberán-Soler, 2002; Alcaraz-Peraza, 2005 y Vilchis-Peluyera, 2006). Un factor determinante en los casos mencionados es el nivel de expresión de mRNA en la línea germinal o en etapas tempranas de las células embrionarias. Genes con una mayor cantidad de mRNAs transcritos en células germinales son más propensos a ser retroprocesados que los genes menos transcritos. Sólo los mRNA que sean retroprocesados en las etapas tempranas del desarrollo pueden ser heredados y llegar a fijarse en la población (Zhang *et al*, 2003). Además muchos de los retropseudogenes encontrados provienen de genes expresados en varios estados de espermatogénesis (Wine *et al*, 1997).

Otra característica observada en los genes de proteínas ribosomales es la existencia de una correlación negativa entre la abundancia de retropseudogenes y el contenido de GC (Guanina-Citosina) del gen original (Gonçalves *et al*, 2000). Zang y colaboradores en el 2003, mencionan que los genes relativamente pobres en GC tienen más retros que los genes ricos en GC y esto posiblemente fuese debido a que los retros originados por secuencias pobres tienen tasas menores de decaimiento que los retros con altos contenidos de GC. Otra explicación para niveles bajos de GC en los genes que originan retros es dada por Korenberg y Rykowski en 1988, los cuales mencionan que las secuencias LINE tienen bajos niveles de GC, por tanto la reverso transcriptasa actuará en mayor medida sobre los mRNA de genes con bajo contenido de GC, es decir, que presenten un mayor parecido con las secuencias LINE.

Otro factor establecido es el tamaño de la secuencia (Gonçalves *et al*, 2000), puesto que en mensajeros grandes existen mayores posibilidades que la transcriptasa reversa se pare antes de terminar, generándose fragmentos que pueden perderse fácilmente por inserciones y/o deleciones. Además, los retropseudogenes grandes ya fijados son más propensos a ser truncados que los cortos (Zhang *et al*, 2003).

Por último Gonçalves y colaboradores reportaron que los genes altamente conservados o que pertenecen a familias génicas altamente conservadas son más propensos a generar retropseudogenes (Gonçalves *et al*, 2000).

En ocasiones las retrocopias pueden transcribirse si se insertan cerca de una región promotora de una secuencia no relacionada e incluso dar origen a nuevos genes.

Este proceso así como el rearrreglo de exones son de vital importancia en la creación de nuevas estructuras genéticas (Long *et al*, 2003). Un ejemplo de esto es el gen quimérico *jingwei* en *Drosophila teissieri* y *D. yakuba* el cual fue creado cuando un mRNA procesado de alcohol deshidrogenasa fue reversotranscrito y reinsertado dentro del tercer intrón del gen llamado del emperador amarillo (*ymp*, por sus siglas en ingles). Como consecuencias de ello se generó un nuevo gen de alcohol deshidrogenasa por su inserción en una región genómica con un sistema regulatorio preexistente y apropiado. Se cree que dicha enzima presenta una nueva función, aunque ésta es aún desconocida (Wang *et al*, 2000).

Otro ejemplo de retrosecuencias que puede adquirir nuevas funciones es el gen de la preproinsulina I en ratas, el cual se originó por retrotransposición del gen de preproinsulina II (Soares *et al*, 1985).

En el presente trabajo se analizaron las retrosecuencias de la familia de las globinas y la argininosuccinato sintetasa, así que en este apartado se dará una breve explicación de estas proteínas y sus genes.

Globinas

En los vertebrados las moléculas transportadoras de oxígeno son la hemoglobina y mioglobina. La primera se encarga del transporte de oxígeno en la sangre, así como también de dióxido de carbono y de iones hidrógeno; la segunda, por su parte, facilita el manejo de oxígeno en los músculos (Stryer, 1995).

La hemoglobina de los vertebrados está formada por cuatro cadenas polipeptídicas, dos de un tipo y dos de otro. En el estadio adulto se producen dos tipos de hemoglobina, A y A₂. La A consta de dos cadenas alfa (α) y dos beta (β), mientras la A₂ está conformada por dos cadenas alfa (α) y dos delta (δ); esta última sólo constituye aproximadamente el 2% del total de la hemoglobina. Durante el estadio embrionario se producen otros tipos de hemoglobinas. Después de la concepción, el embrión sintetiza una hemoglobina conformada por dos cadenas dseta (ζ) y dos épsilon (ϵ), equivalentes a α y β respectivamente. Posteriormente durante el desarrollo fetal las cadenas ζ son

reemplazadas por α y las ϵ por γ , esta hemoglobina es llamada F (Stryer, 1995) (ver Figura 2).

Esta variación en los tipos de hemoglobinas en los diferentes estadios de vida se deben a que cada una de ellas tiene diferentes capacidades de transporte de O_2 , CO_2 y H^+ y corresponde a las necesidades de transporte de dichos de O_2 en estas diferentes etapas (Stryer, 1995)

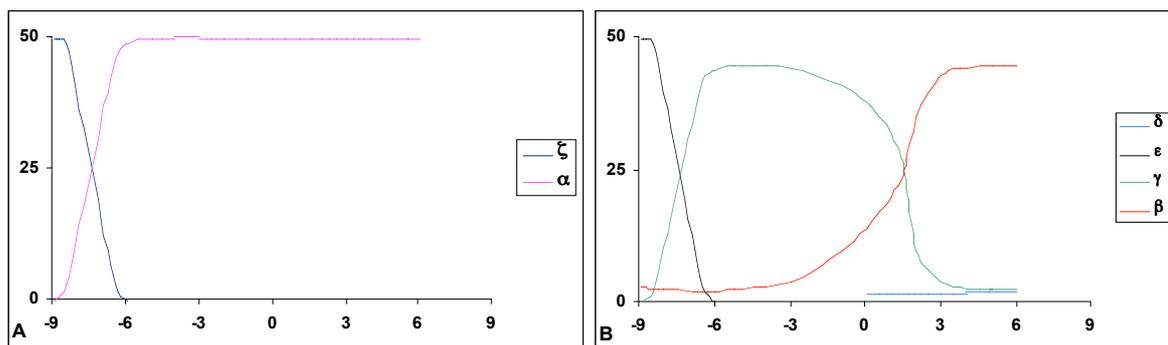


Figura 2. Expresión de los genes de la hemoglobina en los diferentes estadios del desarrollo humano. El eje de las X representa el porcentaje total de cadenas de globina mientras en el otro eje se muestra la edad del individuo en meses, el punto cero marca el nacimiento. A) genes α y ζ B) genes β , γ , δ , ϵ (Tomada de Stryer, 1995)

Globinas como una familia multigénica

La familia de genes que originan las diferentes globinas es un modelo evolutivo extensamente estudiado puesto que es un caso muy evidente de las llamadas familias multigénicas. Estas se originan por duplicaciones completas de un gen o grupo de genes relacionados. A estas familias pertenecen todos los genes, funcionales o no, originados por dichas duplicaciones, los cuales pueden estar en el mismo o en diferentes cromosomas (Li y Graur, 2000).

De manera general, la llamada superfamilia de las globinas está integrada por las recientemente descritas neuroglobina y citoglobina (Pesce *et al*, 2002), además, de la mioglobina y las familias de hemoglobinas α y β (ver Figura 3). Esta familia engloba las diferentes posibilidades que tienen los genes al duplicarse, los cuales son: 1) retención de la función original (como α_1 y α_2), 2) adquisición de nuevas funciones (divergencia de neuroglobina, citoglobina, mioglobina y las hemoglobinas) y 3) pérdida de la función (generándose de esta manera pseudogenes) (Li y Graur, 2000).

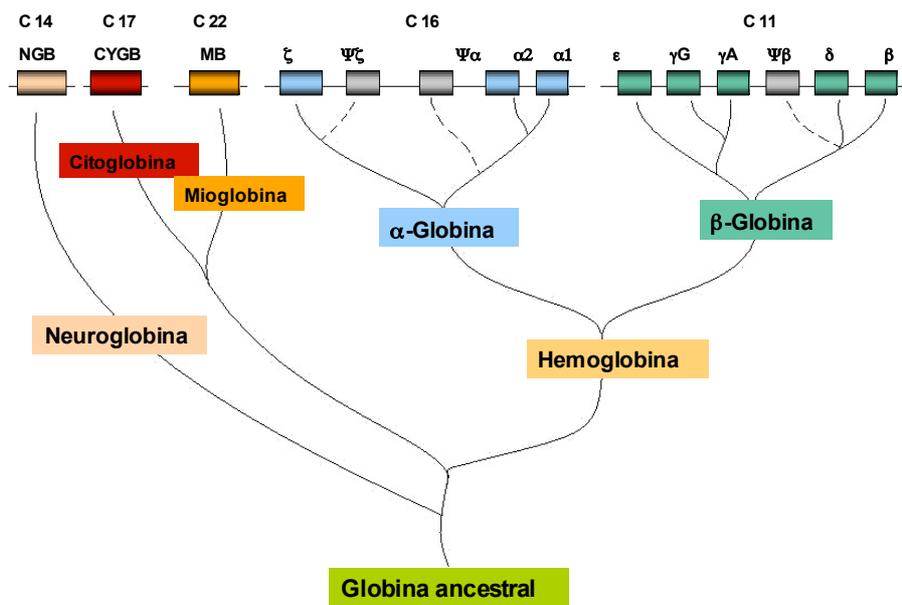


Figura 3. Evolución de la superfamilia de las globinas a partir de una globina ancestral. Las líneas punteadas relacionan a los pseudogenes con el gen que les dio origen. (Tomada de Pesce *et al*, 2002).

Argininosuccinato sintetasa

La Argininosuccinato sintetasa (ASS, L-citrulina, L-aspartato ligasa, EC. 6.3.4.5) es una enzima que se encarga de catalizar la condensación reversible ATP-dependiente de citrulina y aspartato para formar argininosuccinato, el precursor inmediato de la arginina. Se trata de una enzima expresada en varios tipos celulares como células de tejido epitelial, de hígado, riñón, intestino, entre otros (Husson *et al*, 2003). Esta enzima participa en tres importantes procesos, los cuales son: 1) Ciclo de la urea, 2) Síntesis de arginina a partir de citrulina y 3) Síntesis de óxido nítrico (NO) (ver Figura 4).

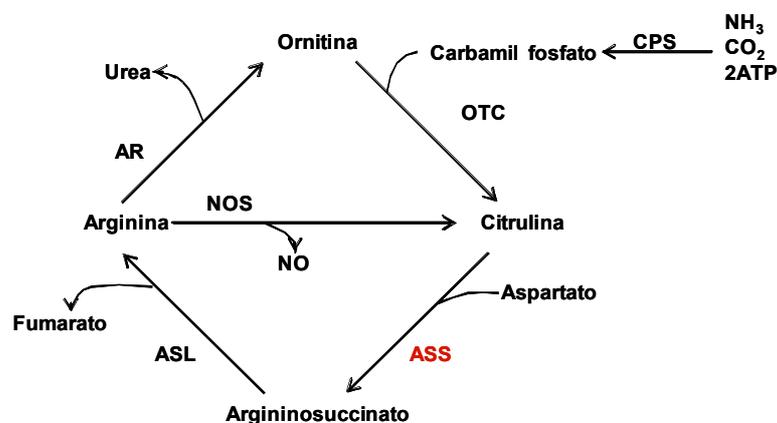


Figura 4. Esquema general de los ciclos en los que participa la argininosuccinato sintetasa (Tomada de Mori *et al*, 1998).

El gen de la ASS en el humano tiene una longitud total de 56.36 kb, con 15 exones. La longitud de su transcrito es de 1572 pares de bases (pb) y el polipéptido resultante presenta 412 residuos de aminoácidos. La enzima ASS, por su parte, es un homotetrámero, en el que cada subunidad esta compuesta de 412 residuos de aminoácidos. Dicha proteína está altamente conservada entre el humano, rata, ratón, bovinos y levaduras (Lemke y Howell, 2002), como se pudo constatar al obtener el porcentaje de identidad entre ellas (Tabla 1). Por otro lado, se han identificado los factores que intervienen en la regulación de la ASS. Estos incluyen hormonas, nutrientes y estímulos proinflamatorios, los cuales actúan a nivel transcripcional (Husson *et al*, 2003).

	Rata	Ratón	Humano	Bovino	Levadura
Rata	100				
Ratón	98.3	100			
Humano	97.6	96.8	100		
Bovino	96.1	95.6	96.1	100	
Levadura	50.4	50.4	50.4	50.6	100

Tabla 1. Porcentajes de identidad del polipéptido de ASS entre los diferentes grupos.

Ciclo de la urea

La serie de reacciones del ciclo de la urea fue propuesta por Krebs y Heseleit en 1932, siendo esta la primera vía metabólica cíclica en descubrirse. La síntesis de arginina y su conversión a ornitina y urea es la principal vía de eliminación de desechos nitrogenados en los mamíferos, dichos desechos son generados principalmente por el catabolismo proteico (Teixeira *et al*, 2003).

Este ciclo (ver Figura 4) requiere la participación de seis enzimas: carbamil fosfato sintetasa I (CPS 1), N-acetilglutamato sintetasa (NAGS), ornitina carbamil transferasa (OTC), Argininosuccinato sintetasa (ASS) y Argininosuccinato liasa (ASL). La actuación final de la argininasas (AR) permite liberar de la arginina la molécula de urea y regenerar la ornitina para continuar el ciclo. En el adulto, el ciclo completo se lleva a cabo en el hígado donde las enzimas están co-localizadas, sobre todo en los hepatocitos periportales (Teixeira *et al*, 2003).

Síntesis de óxido nítrico

El óxido nítrico (NO) es el principal mensajero en actividades como la dilatación de vasos sanguíneos, así como también es un señalizador de la función inmune, específicamente en la activación de macrófagos y funciona como neurotransmisor en el cerebro y sistema nervioso periférico (Knowles y Moncada, 1994).

El NO se sintetiza a partir de arginina por la óxido nítrico sintetasa (NOS), uno de los subproductos de esta síntesis es la citrulina, que puede ser reciclada a arginina por acción de la ASS y la argininosuccinato liasa (ASL), (ver Figura 4). La producción de NO intracelular esta limitada por la disponibilidad de arginina. La arginina puede ser tomada del torrente sanguíneo, de la degradación intercelular de proteínas o por síntesis endógena. El principal productor de arginina es el hígado pero existen otros tejidos en donde se han observado niveles bajos de ASS y ASL, las cuales como ya se mencionó, permiten la síntesis de arginina a partir de citrulina (Mori *et al*, 1998).

Síntesis de arginina

La arginina no es un aminoácido esencial para humanos y roedores ya que éste es sintetizado por células epiteliales del intestino delgado en colaboración con células del túbulo proximal del riñón (Morris, 2002).

Aproximadamente un 28% de la glutamina (un producto final del catabolismo de los aminoácidos en el músculo), vertida en el torrente circulatorio es captada por el intestino y liberada posteriormente a la circulación como citrulina. Una mínima cantidad de esta se queda en el hígado y otros tejidos para una posterior utilización y la mayor parte (83%) se convierte a arginina en el riñón. Las enzimas que participan en la síntesis de arginina incluyen una parte de las enzimas del ciclo de la urea, como son la ASS y la ASL (ver Figura 4) (Teixeira *et al*, 2003).

Gran parte de la arginina sintetizada en el túbulo proximal entra en la circulación sistémica. Una pequeña cantidad se utiliza y/o metaboliza hasta urea en el propio riñón debido a la relativamente baja actividad argininasa de este órgano (Teixeira *et al*, 2003).

I. Justificación

El análisis de los nucleótidos de familias génicas incluidos los pseudogenes y retropseudogenes puede ser una herramienta valiosa para el estudio filogenético que indague en la evolución de los genomas. A partir de dichos análisis se pueden efectuar inferencias evolutivas inter o intra especies (Harrison, 1998).

Los retropseudogenes así como los pseudogenes son modelos ideales para el estudio de las tasas de mutaciones espontáneas. Además, al compararlos con su secuencia original, se puede obtener información acerca de los mRNAs expresados sólo en condiciones particulares, mRNA extintos ó mRNA difícil de clonar y analizar (D'Errico *et al*, 2004).

En general, la importancia de la caracterización de los retropseudogenes, especialmente en el genoma humano incluye las siguientes áreas:

- 1) Por su gran similitud con el gen original funcional los pseudogenes interfieren en las pruebas de PCR o hibridación *in situ* (Hurteau y Spivack, 2002). Al identificarlos, los investigadores pueden usarlos para diseñar secuencias de prueba específicas para el gen original o usarlas como referencia para interpretar los resultados experimentales correctamente.
- 2) Los pseudogenes proporcionan un registro molecular de la dinámica y evolución de los genomas así como de la tasa de sustitución de nucleótidos (Graur *et al*, 1989; Zhang y Gerstein, 2003). De manera general se puede decir que proveen un registro molecular de la evolución de los organismos (Zhang *et al*, 2003).
- 3) Los pseudogenes inducen a errores en las notaciones de las bases de datos. Al tener conocimiento de su ubicación específica se podrá evitar confundirlos con sus contrapartes funcionales en dichas bases (Zhang *et al*, 2003).

Sabiendo que las globinas han sido un grupo altamente estudiado desde el punto de vista evolutivo, se decidió indagar la presencia de retropseudogenes en esta familia.

Los estudios anteriores muestran que los genes que evolucionaron a partir de la globina ancestral y sus pseudogenes asociados son el resultado, únicamente, de procesos duplicativos sin la presencia hasta el momento observada de procesos de retrotranscripción (Li y Graur, 2000).

Posteriormente, se decidió abordar el gen de la ASS por presentar características que han sido descritas como ideales para este tipo de estudios, como son: ser un gen de copia única, lo cual impide que hayan confusiones por las diferencias intrínsecas entre las copias en el genoma, además de ser un gen constitutivo y altamente expresado (Barberán-Soler, 2002 y Alcaraz-Peraza, 2005).

II. Antecedentes

En nuestro grupo de trabajo se han analizado los retroseudogenes asociados a las proteínas de choque térmico HSP10 y HSP60 en el genoma humano así como en el genoma del ratón. Para el genoma humano se identificaron un total de 26 retroseudogenes para la HSP10 y 14 para la HSP60 (Vilchis-Peluyera, 2006), por su parte el genoma del ratón presentó 27 retroseudogenes asociados a la HSP10 con edades de 7 a 67 ma y 40 para HSP60 que oscilan entre los 4 y 68 ma (Alcaraz-Peraza, 2005). Los dos genomas muestran patrones de expresión diferentes para estas enzimas resultado de sus historias evolutivas divergentes desde hace 85 ma.

Estos trabajos ilustran la importancia de los retroseudogenes en el análisis evolutivo de los genomas, incluyendo casos como la identificación de secuencias retroprocesadas con adquisición de nuevas funciones como es el caso en el genoma del ratón del gen del factor de embarazo temprano (EPF), el cual se es esencial para la iniciación y mantenimiento del embarazo en ratones. Dicho gen se originó por retrotransposición del gen de HSP10 en la línea de los roedores adquiriendo una función diferente de la realizada por el gen original (Fletcher *et al*, 2001).

Para el caso particular de la ASS, se han descrito un total de 14 pseudogenes procesados asociados a ésta, de los cuales los presentes en los cromosomas X y Y podrían representar un solo evento de retrotransposición seguido de una duplicación o un proceso de recombinación (Freytag *et al*, 1984a). Gerstein por su parte predice sólo 10 retroseudogenes para esta enzima distribuidos en 8 cromosomas (www.pseudogen.org). Sin embargo, se desconoce si en verdad estos son todos los pseudogenes asociados a dicho gen o existen otros que las técnicas utilizadas por los autores anteriores no permitieron descubrir, además de la falta de información específica sobre cada una de las retrocopias y su relación con el gen original.

III. Objetivos

Objetivo General

Efectuar el análisis de los retropseudogenes asociados al gen de la enzima Argininosuccinato sintetasa, en el genoma humano para identificar los posibles patrones de expresión en el pasado así como su desarrollo dentro de un entorno metabólico y evolutivo.

Objetivos Específicos

- Tratar de identificar posible retropseudogenes asociados a la familia de las Hemoglobinas en el genoma humano. Como hipótesis de trabajo, no deben de encontrarse retropseudogenes, ya que la hemoglobina sólo se expresa en eritoblastos, no habiendo expresión de estas en las células de la línea germinal.
- Identificar los retropseudogenes asociados a la Argininosuccinato sintetasa en el genoma humano.
- Mapear y caracterizar los retropseudogenes encontrados.
- Analizar la identidad de las secuencias con respecto al cDNA original.
- Realizar un análisis de los retropseudogenes recuperados para determinar posibles eventos de duplicación.
- Realizar la datación de las secuencias y analizar los posibles panoramas de patrones de expresión de la ASS en el genoma humano a lo largo del tiempo.
- Realizar una búsqueda de los retropseudogenes de las enzimas asociadas a la ASS en los diferentes ciclos en los que participa.
- Analizar los retropseudogenes de ASS en chimpancé y ratón y su comparación contra los encontrados en humano.

IV. Metodología

Obtención y caracterización de las retrosecuencias

Todas las secuencias de genes, cDNA, proteínas y secuencias de posibles retropseudogenes aquí presentes se tomaron de dos bases de datos públicas, ENSEMBL del Centro de Secuenciación Sanger (www.ensembl.org), así como del Centro Nacional de Información Biotecnológica (NCBI, www.ncbi.nlm.nih.gov), en el subsitio de cada una de las especies analizadas –humano, chimpancé y ratón –.

Para la búsqueda de las retrosecuencias se utilizó como sonda bioinformática al cDNA de la enzima a estudiar en humano, chimpancé ó ratón, según fuera el caso, esto con el fin de evitar las concordancias debidas a los intrones, partiendo de la premisa que los retropseudogenes son secuencias originadas a partir de un mRNA maduro sin intrones. El programa empleado para este propósito fue el programa de búsqueda heurístico Blast (Altschul *et al*, 1990), en la subrutina destinada para nucleótidos Blast-n del ENSEMBL. Este programa se basa en un método heurístico ya que consta de un algoritmo que basado en pruebas y aproximaciones puede llegar a encontrar secuencias concordantes, además cuenta con diversos parámetros que permiten hacer más o menos sensible la búsqueda de secuencias. Uno de estos parámetros es el valor “e”, el cual determina la proporción de secuencias concordantes que es posible encontrar debido únicamente al azar, con un valor de $e = 10$ se puede esperar que 1 de cada 10 secuencias arrojadas por el programa sea debida al azar. Para el presente trabajo se utilizaron diferentes valores de e que fueron de 10 a 0.001, en la mayoría de los caso para verificar que no se estuvieran omitiendo secuencias importantes al poner valores de e muy estrictos, pero los datos aquí presentados son el resultado de un valor de $e = 0.01$.

Otros parámetros a tener en cuenta son la longitud de palabra (w), es decir el número mínimo de nucleótidos que deben ser concordantes para que el programa tome a la secuencia como un “hit”; su valor fue $w = 15$ y los filtros de secuencias de baja complejidad, los cuales como su nombre lo indica impiden que haya concordancias debidas a secuencias de baja complejidad, habiendo ajustado los parámetros anteriores se procede a la búsqueda. El programa despliega los resultados de la búsqueda de manera gráfica y arroja las secuencias concordantes a la secuencia problema; debido a

que este programa se basa en similitudes locales los “hits” de una región dada pueden tratarse de una sola secuencia que fue fragmentada por dicho programa, por tanto cada hit no es necesariamente una secuencia recuperada y para obtener la secuencia real se tienen que empalmar varios de estos hits.

Recuperadas las secuencias candidatas de posibles retropseudogenes, se toman aquellas secuencias que el programa nos informa que se encuentran en dirección contraria, es decir que fueron integradas al genoma en dirección 3' 5' y se obtiene su complemento inverso, este paso se lleva a cabo con ayuda del programa BioEdit 7.0.4.1. Este programa también fue empleado para la visualización de todas las secuencias debido a su interfase gráfica altamente esquemática y fácil de entender, así como para editar las secuencias y los alineamientos resultantes entre estas.

Teniendo todas las secuencias en la misma dirección se procedió a realizar los alineamientos de cada una de ellas contra el cDNA de la enzima a analizar utilizándose para ello el programa Clustal W (Higgins *et al*, 1994) que se encuentra en el sitio Web del Instituto Europeo de Bioinformática (EBI) (<http://www.ebi.ac.uk/clustalw/index.html>). Este programa se empleó para todos los alineamientos pareados y múltiples de nucleótidos o polipéptidos realizados en el presente trabajo. Contando con los alineamientos pareados, se analizó cada uno de ellos para determinar si se trataban de retrosecuencia bajo los siguientes parámetros: su similitud con el cDNA original, la presencia de una cola de poliA o vestigios de esta y la ausencia de intrones o inserciones no mayores a 60 pb (tamaño mínimo de un intrón) en los sitios de empalme de los exones.

Para cuantificar el parecido entre las secuencias se obtuvieron los porcentajes de identidad relativa y absoluta de la secuencia candidata contra el cDNA con ayuda del programa MEGA 3 versión β (Kumar *et al*, 2004) donde se cuantificaron las concordancias dentro de las secuencias (match) así como las discordancias entre ellas (mismatch). Como se puede observar en la ecuación 1 (Ec 1) la identidad relativa no tiene en cuenta el tamaño de las secuencias, lo cual resulta muy útil cuando se trabaja con secuencias de diferentes tamaños como es el caso de los retropseudogenes donde la reverso transcriptasa se detuvo antes de completar la secuencia. Por su parte la identidad

absoluta (Ec 2), sí tiene en cuenta la longitud de las secuencias y permite evaluar la identidad global entre ellas.

$$\%ID_{relativa} = \left(\frac{Match}{Match + Mismatch} \right) \times 100 \dots\dots\dots (Ec 1)$$

$$\%ID_{absoluta} = \left(\frac{Match}{L} \right) \times 100 \dots\dots\dots (Ec 2)$$

Donde:

Match: los nucleótidos o aminoácidos concordantes

Mismatch: los nucleótidos o aminoácidos no concordantes

L: tamaño de las secuencias

Para obtener el cDNA de ASS que dio origen a las retrosecuencias presentes en el genoma del humano y chimpancé se tomaron todas las isoformas reportadas en las bases de datos del ENSEMBL para esta enzima y se alinearon contra las retrosecuencias determinando si existían regiones con mayor identidad entre las retrosecuencia y la isoforma, para posteriormente hacer el empalme de aquellos exones que obtuvieron las mayores identidades contra las retrocopias, además de mapear el cDNA nuevo contra el gen y obtener los porcentajes de identidad relativa nuevos.

Para el presente estudio fue importante descartar aquellas secuencias que se hayan originado por un proceso de duplicación a partir de un retropseudogen puesto que dichas secuencias generan errores al estimar los tiempos de generación de las retrosecuencias y los patrones de expresión de las enzimas. El análisis de posibles eventos de duplicación se hizo mediante la construcción de un árbol basado en el método de Neighbor-Joining, el cual se fundamenta en el principio de mínima evolución (Saitou y Nei, 1987) con todas las retrosecuencias más 200 pb arriba y abajo en cada una, además del cDNA original, con un bootstrap 550 repeticiones. La matriz de distancias se calculó con el método de Kimura de 2-Parámetros (Kimura, 1980); dicho método pondera de diferente manera los procesos de transición y transversión entre nucleótidos.

Se debe destacar que este árbol no aporta información acerca de la filogenia de los retroseudogenes puesto que se parte de la premisa que cada uno de ellos fue un evento independiente de retrotranscripción, pero nos permite saber si existen retrocopias que sean más parecidas entre si que con respecto al cDNA original, este dato aporta el primer indicio de un posible evento de duplicación entre estas, lo cual posteriormente debe corroborarse mediante el análisis de la identidad entre las secuencias y de sus secuencias flanqueantes.

Datación de las secuencias

Para la datación de las secuencias se utilizó el modelo de Jukes-Cantor para el cálculo de las distancias entre las retrosecuencias y la secuencia funcional, este modelo considera que no existe distinción entre las transiciones y las transversiones, por lo cual se le denomina modelo de un solo parámetro (Ec 3).

$$K = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p \right) \dots\dots\dots(\text{Ec 3})$$

Donde K es el número de sustituciones por sitio a partir del tiempo de divergencia y p es el grado de divergencia entre las secuencias a comparar y está dado por el número de cambios existentes entre las secuencias dividido por la longitud total de estas (Ec 4) (Li y Graur, 2000).

$$p = \frac{n}{N} \dots\dots\dots(\text{Ec 4})$$

Para obtener el tiempo de divergencia conociendo las distancias de cada secuencia con respecto a la secuencia funcional, se debe partir del concepto de reloj molecular, es decir, se debe obtener una tasa constante de cambio para dicho pseudogen que nos permita ubicarlo en una escala temporal pensando que las dos secuencias, tanto la del pseudogen como la secuencia actual han divergido de la secuencia original (Ver Figura 5).

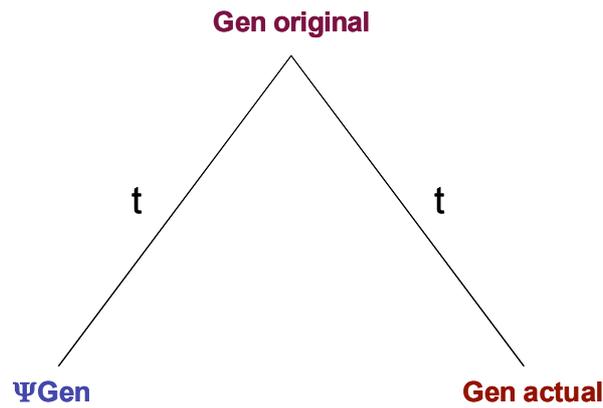


Figura 5. Modelo de divergencia del gen original al pseudogen y al gen actual. Donde t es el tiempo de divergencia.

Diversos autores han calculado tasas de cambio para secuencias funcionales y no funcionales utilizando los datos de las secuencias conocidas. En el presente trabajo se manejará la tasa de cambio propuesta por Li para pseudogenes ($r = 3.71 \times 10^{-9}$ sustituciones/sitio/año) dado que se podría esperar que al no estar sujetos a presiones de selección tanto los pseudogenes como los retropseudogenes tuvieran tasas de cambio parecidas (Li y Graur, 2000), además es difícil proponer tasas de cambio para retropseudogenes debido principalmente a su falta de estudio. Partiendo de la ecuación de tasa de cambio:

$$r = \frac{K}{2T} \dots\dots\dots(\text{Ec } 7)$$

Despejando el tiempo de divergencia T de la ecuación 7:

$$T = \frac{K}{2r} \dots\dots\dots(\text{Ec } 8)$$

Potencial codificante

Para determinar el posible potencial codificante de las retrosecuencias se utilizaron los programas GenScan y Artemis v7 (Rutherford *et al*, 2000). El programa GenScan despliega diversos parámetros, como el porcentaje de CG, además de predecir el polipéptido de mayor tamaño en los 6 marcos de lectura.

Por su parte el programa Artemis v7 despliega gráficamente y genera la secuencia de los ORFs de mayor tamaño en los 6 diferentes marcos de lectura, además de mostrar los sitios donde se han originado codones de término prematuros. Como se puede apreciar en la Figura 6, las barras en color son los ORFs mayores de 100 aa para cada uno de los seis marcos de lectura, las líneas verticales muestran codones potenciales de término y en la parte inferior se despliega la secuencia de nucleótidos y sus correspondientes polipéptidos en los seis marcos de lectura.

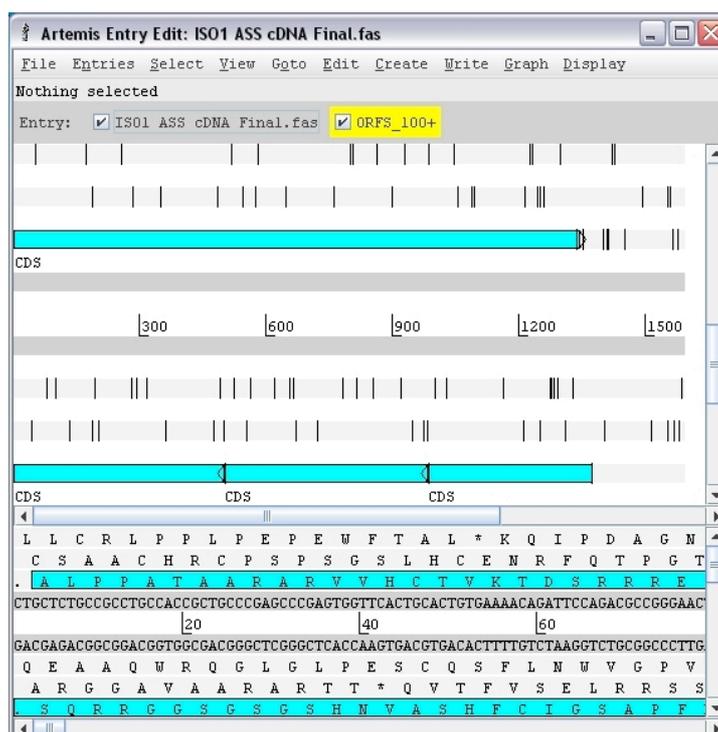


Figura 6. Ventana desplegada por el programa Artemis.

Al comparar los resultados obtenidos con ambos programas se observaron fuertes discrepancias entre ellos, posiblemente debido a que el programa GeneScan predice el polipéptido de mayor tamaño de manera inversa a como se suelen predecir los ORFs, es decir parte de la primera metionina que encuentra y comienza a traducir en sentido 5' a 3' hasta encontrar un codón de término. Debido a esta manera de operar y a que dicho programa no mostraba todos los polipéptidos posibles se descartó para el análisis y solo se tomaron los resultados arrojados por el programa Artemis para ORFs de más de 100 nucleótidos.

Habiendo obtenido los polipéptidos de mayor tamaño en los tres marcos de lectura de 5' a 3' para cada secuencia, se procedió a compararlos con el polipéptido funcional alineándolos con el programa Clustal W para obtener sus porcentajes de identidad.

Posteriormente se llevó a cabo una reconstrucción del polipéptido funcional a partir de la traducción en los 3 marcos de lectura de cada retrosecuencia para poder apreciar cómo las secuencias han sufrido cambios de marco de lectura como consecuencia de las mutaciones, inserciones y/o deleciones. Este polipéptido reconstruido también se comparó con el funcional obteniendo los alineamientos pareados y los porcentajes de identidad.

Análisis de polimorfismos en las secuencias

La secuencia metodológica de esta parte del trabajo consistió en su mayor medida en la observación de los cambios así como en identificar en qué posición (1ª, 2ª ó 3ª) se encontraban dichos cambios; ésto se realizó con la ayuda del programa Artemis v 7 y de los alineamientos múltiples antes realizados.

Otro programa empleado para la realización del presente trabajo fue: Statistica 6.0 para la obtención de histogramas y correlaciones estadísticas.

De la misma manera se recurrió a diferentes bases de datos, como son: a) la base de datos para proteínas EXPASY (<http://ca.expasy.org/>), donde se obtuvo información sobre las secuencias empleadas y participación dentro de las diferentes vías metabólicas y b) la base de datos de en línea de pseudogenes procesados (www.pseudogen.org).

V. Resultados y discusión

a. Búsqueda de retropseudogenes en la superfamilia de las hemoglobinas

Para esta búsqueda se utilizaron los cDNA de β -globina y $\alpha 1$ -globina puesto que estos genes son los principales representantes de cada familia de las hemoglobinas. Con ello se logró una mayor sensibilidad en el análisis. Ambos fueron tomados de las bases de datos del NCBI.

La familia de las β -globinas está presente en el cromosoma 11, en la región 11p15.5. Esta familia se encuentra integrada por las hemoglobinas epsilon (ϵ), gama A ($^A\gamma$), gama G ($^G\gamma$), delta (δ) y por supuesto beta (β). La estructura de sus genes es altamente parecida, teniendo porcentajes de identidad entre ellas de 71.43% de gama A con delta, hasta 98.64% entre gama A y gama G, lo que es indicio de su pasado común. El número de acceso para la familia de la β -globina es NG_000007, con una longitud total de 81.706 kilobases (kb). Mientras el gen de β -globina presenta una longitud de 1.60 kb y el número de su cDNA es NM_000518.4, presentando una longitud de 626pb.

Por su parte la familia de las α -globinas esta presente en el cromosoma 16 en la región 16p13.3, se encuentra conformada por las hemoglobinas alfa 1 ($\alpha 1$), alfa 2 ($\alpha 2$), theta (θ), y dseta (ζ). Al igual que dentro de la familia de las beta, las alfa guardan un gran parecido entre ellas, como resultado de su divergencia a partir de una secuencia ancestral común. La familia de las α -globinas tiene una longitud total de 43.058 kb con número de acceso NG_000006. El gen de la $\alpha 1$ -globina tiene una longitud de 872 pb y su cDNA consta de 576pb, con número de acceso NM_000558.3.

Como se muestra en la Figura 7, con un $e = 0.01$ en el algoritmo de búsqueda Blast-n, la $\alpha 1$ -globina, al igual que la β -globina solo tuvieron concordancia con si misma así como con los miembros de su propia familia. Si se disminuye la sensibilidad por un aumento del parámetro “e” ($e = 0.1, 1.0$ y 10.0), el cual asigna el número de secuencias que pueden ser encontradas por azar, se obtiene un número de hits mayor debido a la aparición de secuencias concordantes por azar. Al depurar los hits

descartando este último tipo de secuencias se obtiene el mismo resultado de $e = 0.01$. Estos resultados muestran una total ausencia de retroseudogenes para las familias de la hemoglobina en el genoma humano.

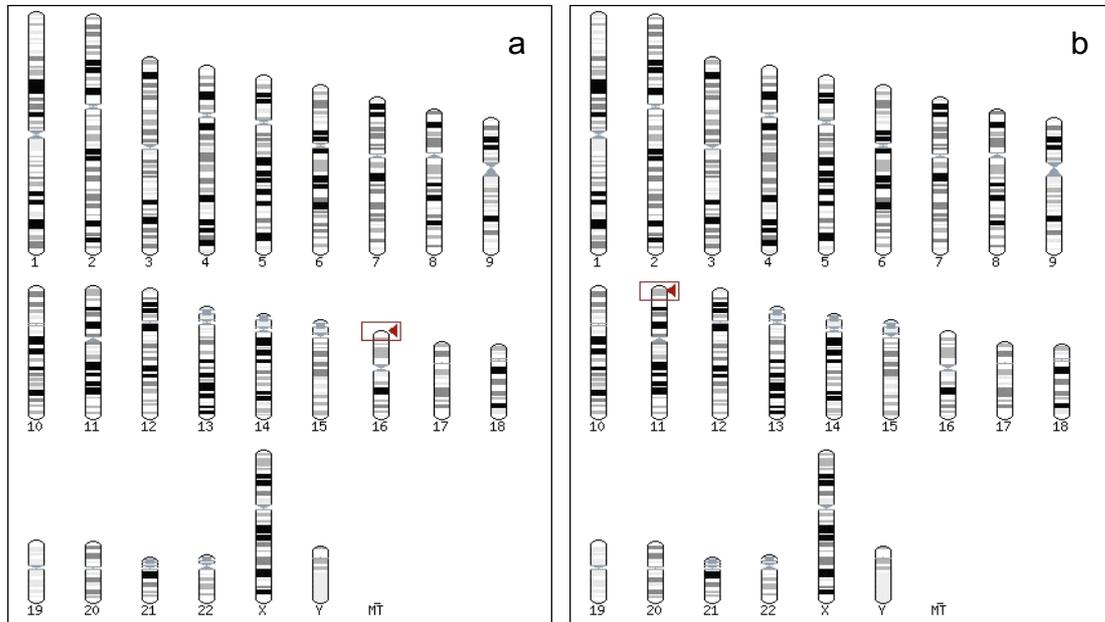


Figura 7. Resultados de Blast-n para: a) α -globina y b) β -globina. En ambos casos el hit representa la ubicación del gen.

La ausencia de retroseudogenes de hemoglobinas en el genoma humano, se puede explicar por su nula expresión en línea germinal, ya que las hemoglobinas son tejido específicas, y se expresan únicamente en los eritoblastos. Los eritoblastos son células altamente especializadas, pues solamente transcriben entre 5% a 10% del material genético, siendo los genes de la globina parte de esta fracción (Telen, 1994). Como ya se ha mencionado, la expresión del gen en línea germinal es un requisito indispensable para que los retroseudogenes puedan pasar a las siguientes generaciones y posteriormente sean fijados en la población.

b. Descripción general y búsqueda de las retrosecuencias para ASS

Según la base de datos del ENSEMBL, el gen correspondiente a la enzima ASS en humano, con número de acceso ENSG00000130707, se encuentra en el cromosoma 9 en la región 9q34.1 y su longitud total es 56.36kb. Este gen presenta 15 exones, a partir de los cuales se genera un mRNA maduro con una longitud de 1572 pb y a su vez éste da lugar a un polipéptido de 412 residuos de aminoácidos (ver Figura 8).

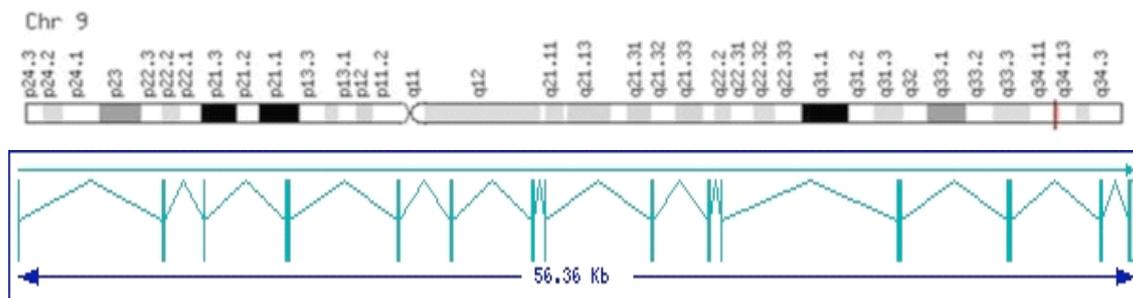


Figura 8. Localización de la ASS dentro del cromosoma 9 y su patrón de exones.

Para hacer la búsqueda de retrocopias se utilizó el cDNA como sonda bioinformática. El Blast-n dio como resultado 42 hits ubicados en 16 posiciones diferentes y distribuidos en 9 cromosomas autosómicos y los dos cromosomas sexuales (ver Figura 9). Del total de hits presentes, se observó que 17 de estos correspondían a alineamientos con el gen original, así que fueron descartados del estudio. Al hacer el empalme de las secuencias correspondientes a los 25 hits restantes se obtuvieron 15 posibles retrocopias. La Tabla 2 muestra algunas de las primeras características presentes en las secuencias encontradas, como son: nombre de la posible retrocopia (tomado de la bases de datos de NCBI); cromosoma en el que está presente; ubicación dentro del cromosoma; dirección original de la secuencia y tamaño de ésta.

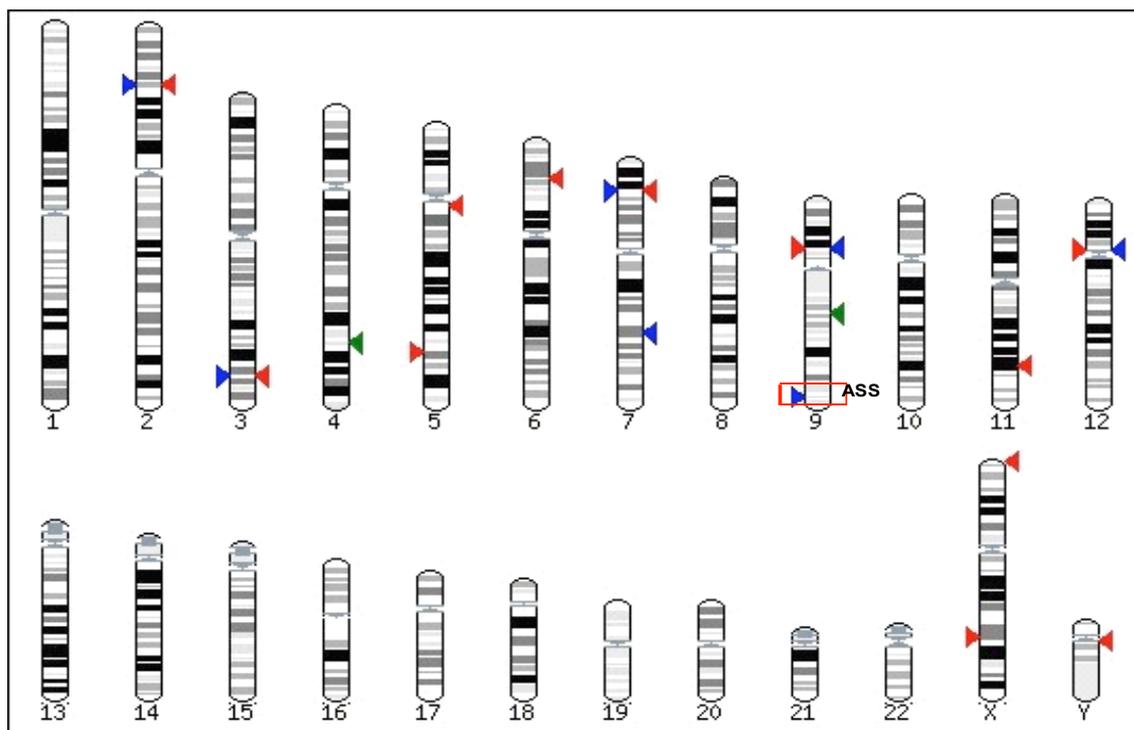


Figura 9. Resultado del Blast-n para la Argininosuccinato sintetasa, el cuadro marca el lugar donde se encuentra el gen.

	Cromosoma	Sitio en Crom	Dir	Tamaño (pb)
ASS	9	9q34.1	+	1592
P01	6	6p22.1	-	1597
P02	2	2p22.1	-	1565
P03	9	9q11-q22	+	1604
P04	X	Xpter-p22	+	1596
P05	X	Xq22-q26	+	1573
P06	Y	Yq11.21	-	1602
P07	3	3q12-qter	-	1590
P08	4	4q31.22	-	897
P09	5	5q11.2	+	1594
P10	5	5q32	+	1594
P11	7	7p15.3	-	1572
P12	9	9p21.1	+	1586
P13	11	11q22.3	+	1564
P14	12	12p11.21	+	1582
P15	7	7q31.1	+	807

Tabla 2. Datos generales de las secuencias recuperadas.

Como primera aproximación para conocer si se trataban de retrocopias, se alineó cada una de las secuencias con el cDNA original y se analizaron las características de las secuencias. Con ello se pudo observar que no existían inserciones mayores de 10pb y no presentaban disrupciones en los sitios de empalme de los exones. Además se determinó la existencia de una señal de poliadenilación para todas las secuencias y la

presencia, aunque no en todos los casos, de vestigios de colas de poliA no mayores a 18pb.

Teniendo los alineamientos pareados se procedió a calcular el porcentaje de identidad relativa de cada secuencia. Se calculó la identidad relativa debido a que ésta no depende del tamaño de las secuencias, pues como se observa en la tabla anterior algunas de las secuencias eran mucho menores al cDNA original. Los porcentajes de identidad van de 85.89 a 91.78% (ver Figura 10), esto nos indica que todas las secuencias obtenidas son homólogas a la ASS. Dado lo anterior se puede afirmar que las 15 secuencias recuperadas son retrocopias provenientes del mRNA de la ASS.

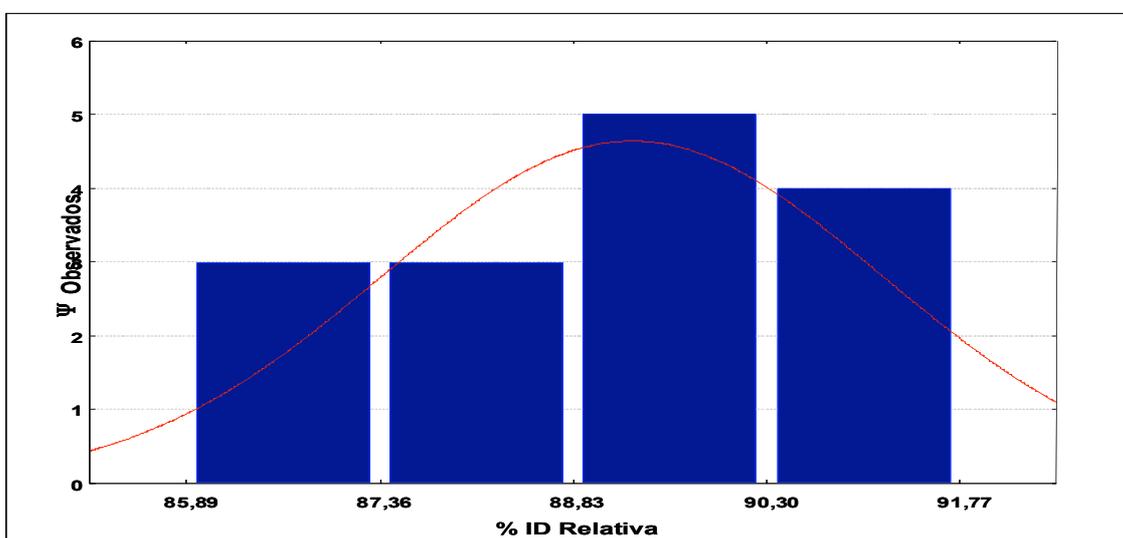


Figura 10. Histograma que muestra la distribución de %Identidad Relativa de las secuencias.

c. Descripción del cDNA

Posteriormente, se observó que al realizar el alineamiento múltiple entre el cDNA y todas las retrocopias, una porción de aproximadamente 75 pb de la región 5' UTR del cDNA no concordaba con la región correspondiente de las retrocopias, pero entre estas últimas existe un alto grado de identidad en dicha región (ver Figura 11). Esto puede deberse a la divergencia de las copias con el gen original, pero es muy poco probable que todas ellas hayan divergido de igual manera. También cabe la posibilidad de que la **isoforma** presente en las bases de datos no sea la que dio origen a las retrocopias, pues se sabe que la ASS presenta splicing alternativo, pudiendo haber más de un cDNA para dicha enzima (Freytag *et al*, 1984b). Lo anterior indica que el mismo gen puede generar cDNA diferentes, debido a empalmes diferentes de los exones e intrones.

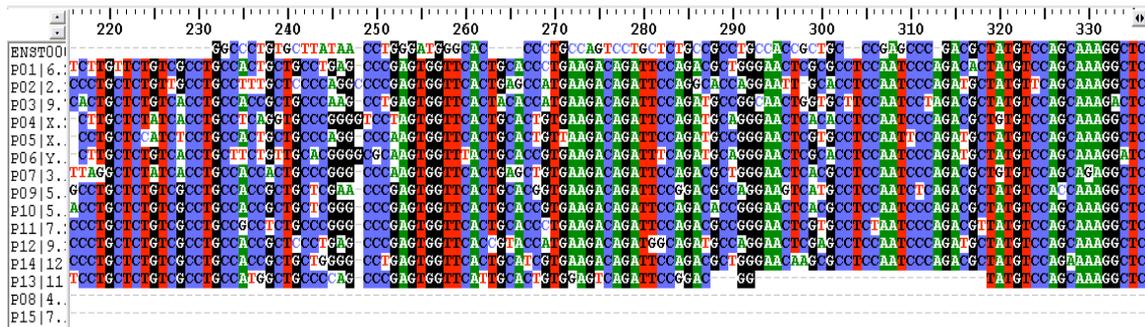


Figura 11. Alineamiento de los retroseudogenes con el cDNA en la región 5' de ASS, propuesto por las bases de datos. Se observa una discordancia de las secuencias con el cDNA, el cual se encuentra en la primera posición.

Revisiones posteriores en las bases de datos mostraron una actualización para ASS donde se presentaban 4 isoformas generadas por splicing alternativo, a diferencia de la versión anterior que sólo mostraba una. De estas isoformas, 2 estaban confirmadas por 15 exones mientras las otras presentaban 16 (ver Figura 12). Las cuatro muestran diversidad en la región 5'UTR, pero a partir del codón de inicio, las secuencias son idénticas. Al comparar sus secuencias de polipéptidos, se observó que codificaban para la misma secuencia, a excepción de la isoforma 3 (Iso3), la cual genera un polipéptido de 488 residuos de aminoácidos, dado que tiene un corrimiento del codón de inicio.

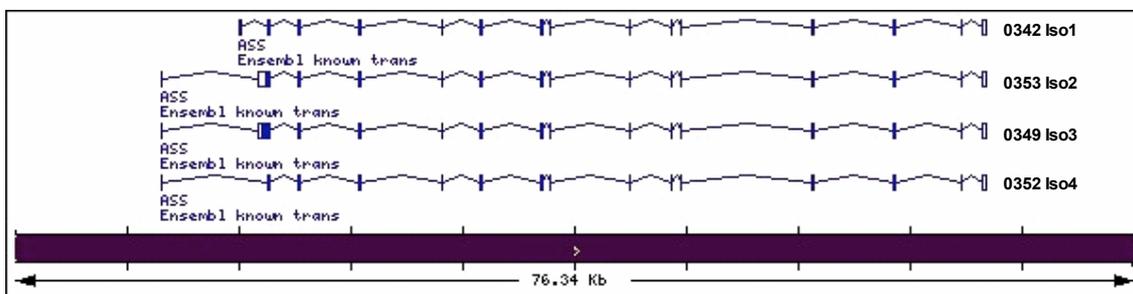


Figura 12. Patrón de exones de las diferentes isoformas de ASS presentes en las bases de datos.

Se procedió a analizarlas comparándolas con las retrocopias y partiendo de dichas isoformas manipular los diferentes exones, generando probables empalmes entre ellos. Al hacer esto último, se encontró que al empalmar el primer exon de la isoforma denominada 3 con el resto de los exones de la isoforma 1 se obtenía un cDNA que al ser alineado con las retrocopias, presentaba un alto grado de identidad con éstas en la región 5'UTR. (Figura 13). Si se compara con la Figura 11 se observa claramente que la isoforma propuesta logra un mejor alineamiento.

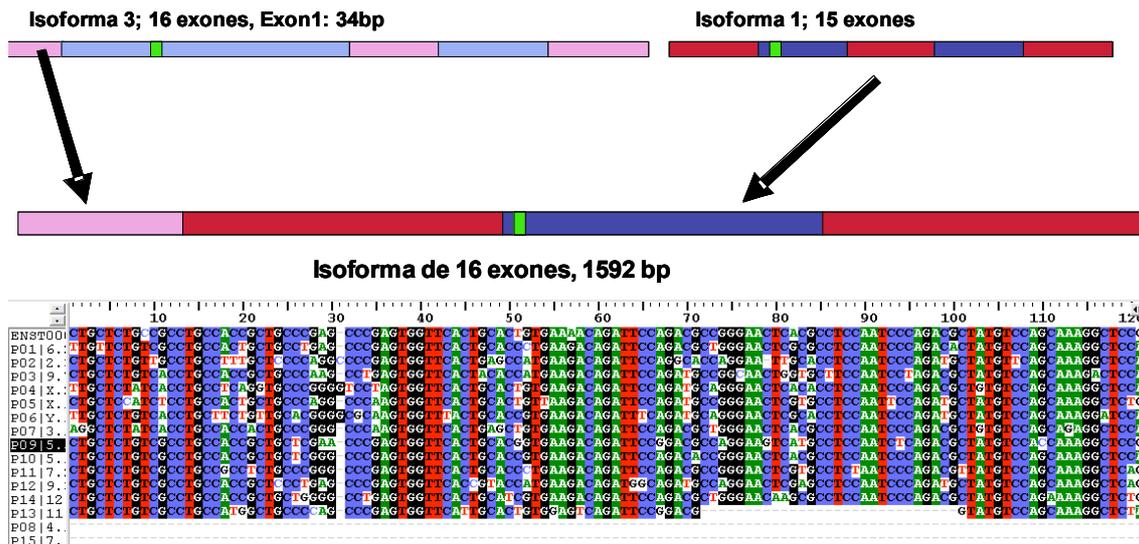


Figura 13. Nueva isoforma para ASS de 16 exones generada por la unión el primer exon de Iso3 con el resto de Iso1, que alinea perfectamente con los retrospudogenes encontrados.

Para verificar dicho parecido se calcularon los porcentajes de identidad de las retrocopias contra la nueva isoforma obteniéndose identidades en el rango de 87.81 a 93.64%, es decir, de hasta dos puntos porcentuales por arriba de las obtenidas con el cDNA anterior. La Figura 14, muestra un comparativo entre los %ID de la secuencia presente en las bases de datos y los obtenidos con el cDNA modificado, donde se aprecia claramente porcentajes mayores para el segundo.

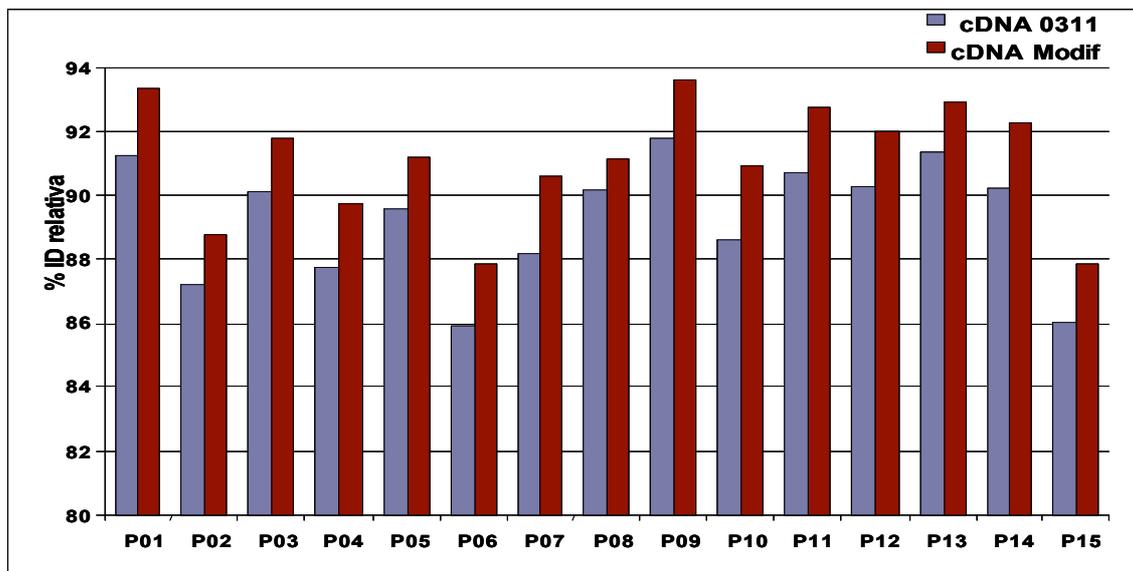


Figura 14. Comparativo del %ID relativa del cDNA de las bases de datos con el cDNA modificado.

El histograma resultante al calcular los porcentajes de identidad relativa del cDNA modificado, muestra también este incremento en los porcentajes, desplazando las

categorías y reorganizándolas (ver Figura 15). Debido a lo anterior, se propuso que esta isoforma derivada fue la que dió origen a las retrocopias existentes en el genoma del humano y se tomó como modelo para el posterior análisis de las retrocopias, descartándose el cDNA anterior (el cDNA completo se encuentra en el Anexo I).

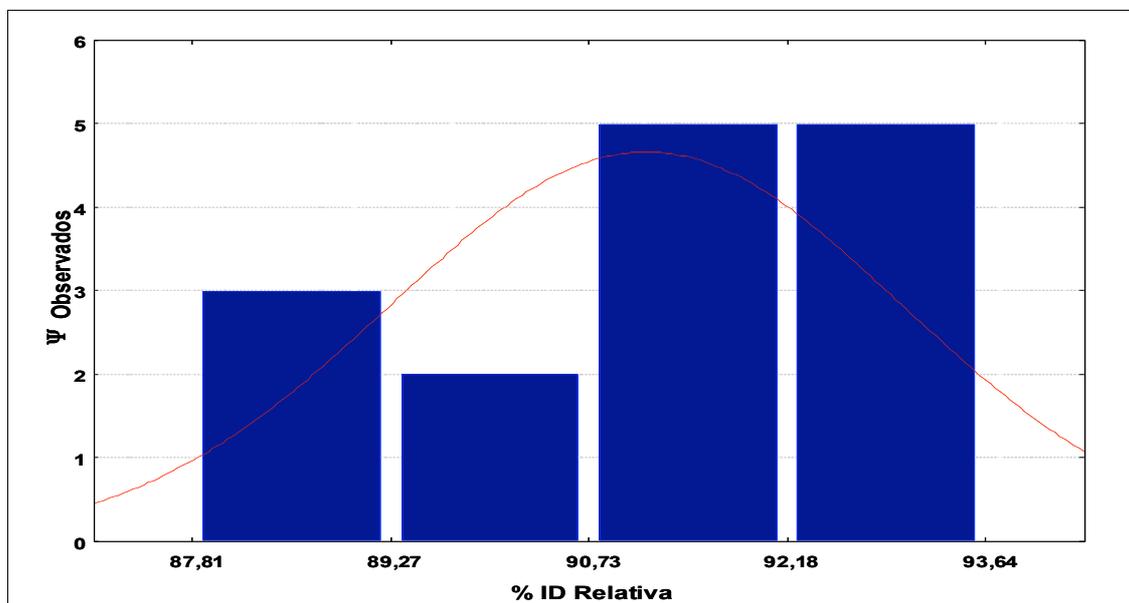


Figura 15. Histograma de distribución de %ID relativa utilizando el cDNA modificado.

Diversos autores han propuesto isoformas para la ASS. Bock y colaboradores en 1983, fueron los primeros en reportar la secuencia del cDNA de la ASS, la cual consistía en 1547 nt más 76 bases de poliA, con una región 5'UTR de 154nt. Freytag y colaboradores en 1984b proponen dos isoformas, la primera (denominada A) con 1578nt y 98nt en la región 5'UTR y la denominada B con una longitud total de 1508nt y 36nt en 5'UTR. Por su parte Pendleton y colaboradores en 2002 hacen la descripción de 3 diferentes isoformas con regiones 5'UTR de 43, 66 y 92nt, encontrando que solo la primera se expresa en el hígado mientras en el tejido endotelial hay expresión de las tres, siendo la primera la predominante.

Nuestro cDNA resultante presenta 16 exones a diferencia del cDNA de las bases de datos que tiene 15; su longitud es de 1593 nucleótidos (nt) contra 1572. El codón inicial se encuentra en el tercer exon, con una región 5'UTR de 102nt, mientras el cDNA de las bases de datos presenta el codón inicial en el segundo exon y su región 5'UTR es de una longitud de 80nt. Cabe destacar que la variabilidad de estas formas únicamente está dada en la región 5'UTR, a partir del codón inicial las secuencias son

idénticas. La Figura 16 muestra el splicing generado para la secuencia de las bases de datos y el del cDNA modificado para la región 5'UTR, a partir del exon que contiene el codón inicial ambas secuencias son idénticas. Cabe mencionar que no existen corrimientos del marco de lectura de la secuencia encontrada, por lo tanto el polipéptido resultante del nuevo cDNA es el que da lugar a la enzima ASS.

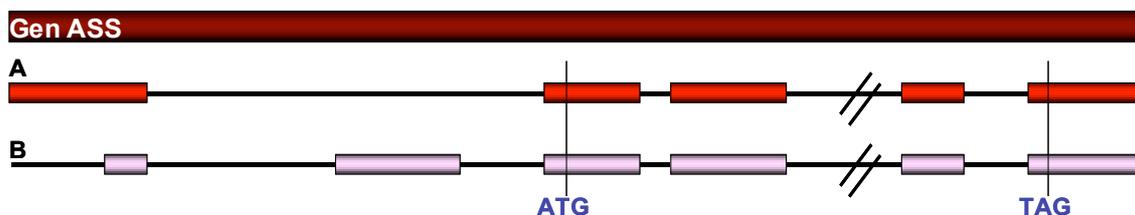


Figura 16. Se ilustra el Splicing para: A) la isoforma que da lugar al cDNA de las bases de datos y B) el que da lugar al cDNA modificado en la región 5'. –Las barras representan los exones, en tanto que las líneas entre ellas representan los intrones– Hay que notar que tanto el codón de inició como el de término se encuentran en la misma posición en ambas secuencias.

Esta variabilidad en la región inicial puede deberse a que en un solo individuo se generen diferentes isoformas tejido-específicas con mecanismos de regulación diferenciales (Pendleton *et al*, 2003). Una región de mayor tamaño puede aletargar la transcripción ó por el contrario, mediante la formación de horquillas, se puede ver favorecida y acelerada la unión del mensajero con el ribosoma. Además es altamente probable que esta diversidad y sus diferentes mecanismos de regulación estén relacionados con los ciclos en los que participa la ASS (ciclo de la urea, síntesis de arginina y síntesis de NO), debido a que, cada uno de ellos presenta requerimientos diferentes de la enzima (Mori *et al*, 1998).

En este sentido, partiendo del hecho que los retropseudogenes son eventos que ocurrieron en gametos o en células tempranas durante el desarrollo ontogenético, es probable que la isoforma propuesta sea la generada en dichos órganos. Reportes como los de Husson *et al*, 2003 indican que existe alta actividad de ASS en testículo de rata tanto en fetos como en organismos adultos. Todo lo anterior tendría que ser demostrado experimentalmente. Otra alternativa podría ser, que la isoforma que dio origen a las retrocopias haya sido la de mayor nivel de expresión en la línea de los homínidos antes de la divergencia del humano (Freytag *et al*, 1984b).

d. Búsqueda de eventos de duplicación

Antes de proceder a la datación de las secuencias y teniendo el cDNA más probable, éste se alineó con todas las retrocopias más 200pb extras arriba y abajo de cada secuencia, realizándose posteriormente un árbol de Neighbor-Joining, con el fin de determinar si existía relación entre ellas o se originaron por diferentes eventos de retrotranscripción (ver Figura 17).

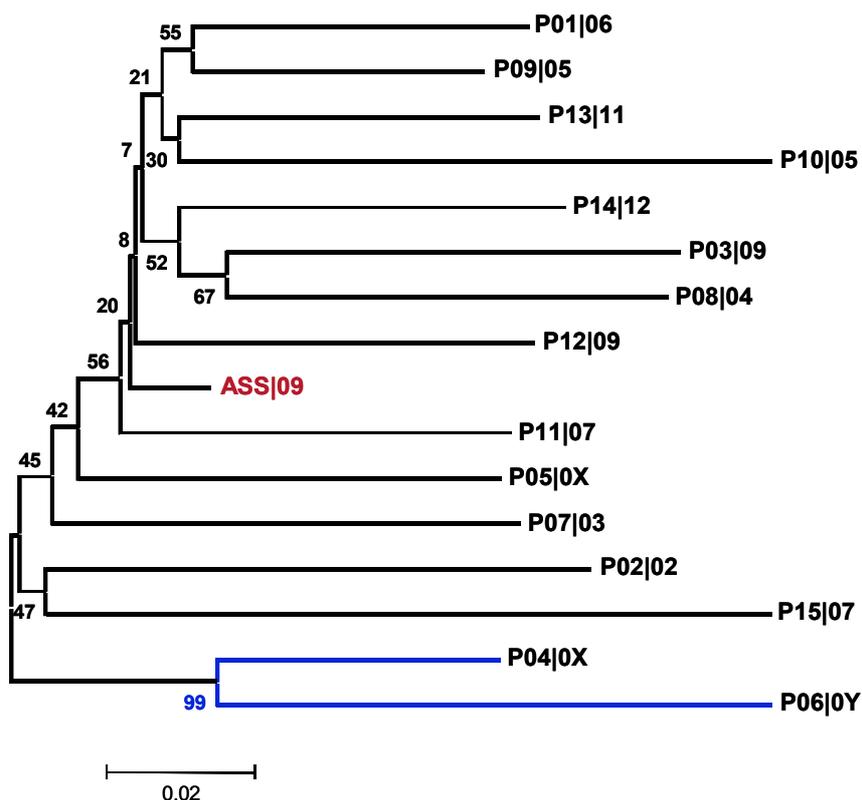


Figura 17. Árbol de Neighbor-Joining que muestra las relaciones entre los retrospseudogenes de la ASS. Se utilizó el modelo de Kimura 2 Parámetros con un Bootstrap de 550 repeticiones. El cDNA original se resalta en rojo y la pareja de retrosecuencias con el bootstrap más alto se resalta en azul.

El análisis mostró que P04 y P06 están altamente relacionadas (con un bootstrap de 99), por lo cual se procedió a alinearlas y obtener la identidad relativa entre ellas utilizando las regiones 5'/3' de 200pb como marcador de posible duplicación. La identidad relativa fue de 90.07%, la cual es mayor a la identidad que presentan cada una con respecto al cDNA – 89.72% para P04 y 87.89% con P06 –. Además, al estudiar las regiones de 200pb flanqueantes se obtuvieron porcentajes de identidad de 88.50 y 84.46% para las secuencias flanqueantes 5' y 3' respectivamente, estos porcentajes no pueden ser debidos al azar. Por lo que, se puede suponer que, se trata de un solo evento de retrotranscripción seguido de un evento de duplicación o recombinación.

Para determinar cual secuencia es la ancestral hay que ahondar un poco en su historia evolutiva. P04 se encuentra en el cromosoma X en el brazo p mientras que P06 se encuentra muy cercana al centrómero en el brazo q del cromosoma Y (ver Figura 9). Estudios de los cromosomas sexuales en humano, han establecido que, a pesar de sus diferencias, estos dos cromosomas comparten una historia evolutiva común, dado que evolucionaron a partir de un par autosómico y aun en la actualidad sigue habiendo relaciones de intercambio de información genética, por recombinación de regiones homólogas, entre ellos. Entre las regiones claramente homólogas en la actualidad tenemos las regiones PAR1 y PAR2, pero también existen homología entre otras regiones que en la actualidad no recombinan (Ross *et al*, 2005).

Estudios como los de Steinemann y Steinemann en 2005, sugieren que el cromosoma Y se encuentra en proceso de degeneración por el silenciamiento masivo de genes al que está sometido. Dicho silenciamiento está dado principalmente por la acumulación de retrotransposones, insertados dentro de genes que en otro momento fueron funcionales. Esto está generando una transformación a heterocromatina de la eucromatina y la supresión de la recombinación con el cromosoma X. Lo anterior, según los autores, es un prerequisite para la determinación de un sistema sexual genéticamente estable, es decir, un sistema X0.

Por tanto, el panorama que se muestra para P04 y P06 es el siguiente: debido a la alta actividad de retrotransposones presente en Y se generó P06 y se insertó en una región rica en elementos altamente repetitivos como lo son los centrómeros. Posiblemente en este momento evolutivo, la recombinación entre X y Y no estuviera tan restringida, por lo que, mediante entrecruzamiento P06 pasa a X, llevándose consigo otras regiones vecinas. Dado que ninguna de las dos secuencias están sujetas a presiones de selección, divergen del cDNA original. Se podría pensar que siendo una derivada de la otra sus identidades con el cDNA tendrían que ser muy similares, pero el porcentaje de identidad indica que la secuencia de Y presenta mayor divergencia con el cDNA, lo que podría indicar que se trata de la más antigua o que la tasa de divergencia de la región donde se encuentra es mayor que la vecindad de P04. No se puede pensar en un solo evento de retrotranscripción o que provengan de un mismo alelo mutante, como lo plantean Freytag y colaboradores en 1984a, pues, en ese caso, las retrocopias serían muy similares pero las secuencias flanqueantes no tendrían por qué serlo.

Basándose en este panorama se descartará a P04 presente en el cromosoma X de los siguientes análisis, debido a que no se trata de un evento de retrotranscripción independiente.

e. Datación de los retroseudogenes

Se procedió a datar las secuencias y se elaboró un histograma agrupándose los datos en cuatro categorías. El histograma resultante muestra que las 14 retrosecuencias se originaron en un periodo comprendido entre 8.92 a 17.85 millones de años (ma) (ver Figura 18), siendo el periodo conformado entre los 8.92 y 13.39 ma el de mayor aparición de retros (10 en total).

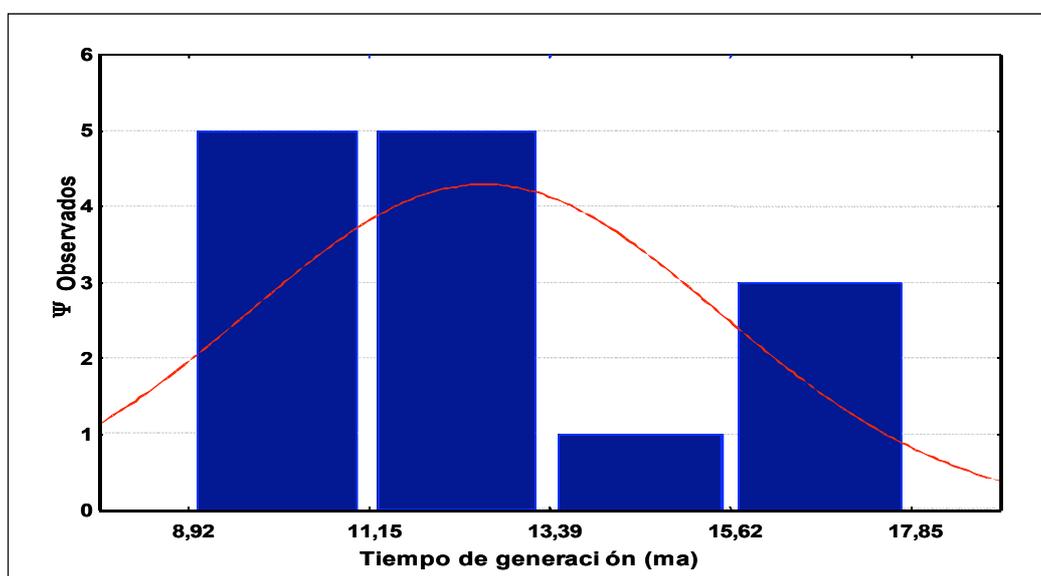


Figura 18. Histograma que muestra la distribución de los tiempos de aparición de las retrosecuencias asociadas a ASS.

La existencia de retrosecuencias en este periodo puede ser evidencia de un aumento significativo en la transcripción de la ASS en gónadas, en comparación con otros periodos históricos, puesto que, al incrementarse la necesidad de una enzima, se dará una mayor transcripción y traducción de su gen, aumentando con ello la probabilidad de generación de retrosecuencias por la actividad de la reversotranscriptasa. Este incremento en la expresión de la ASS en gónadas, pudo deberse a un aumento en los requerimientos del señalizador NO en dichos órganos, lo cual suscitó una elevación en los niveles de la maquinaria de producción de este, que como ya se mencionó incluye a la ASS.

Los ancestros del humano divergieron de los monos del viejo mundo hace aproximadamente 25 ma (ver Figura 19), con los datos obtenidos se puede esperar que los orangutanes y gorilas compartan algunos de los retropseudogenes para la ASS presentes en el genoma humano. En este mismo sentido, podemos esperar también, que el chimpancé comparta con el humano todos los retropseudogenes para la ASS.

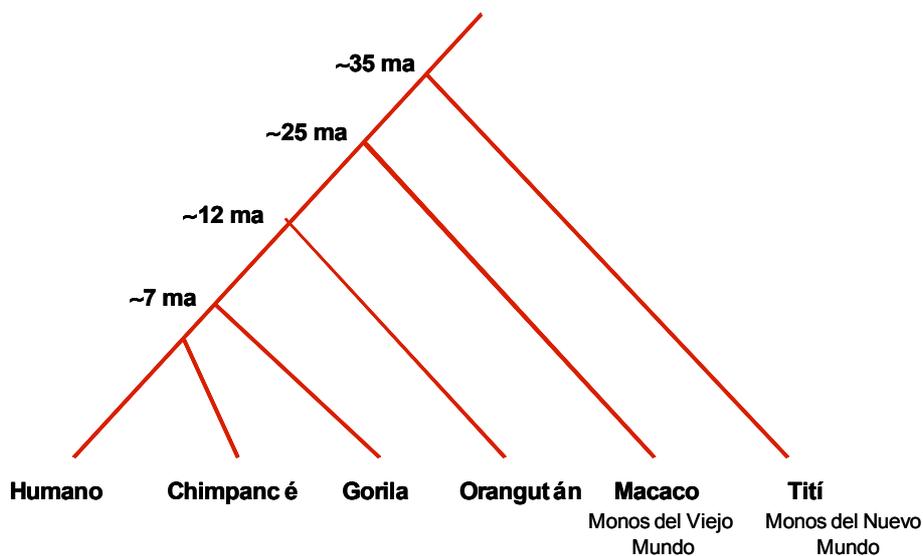


Figura 19. Divergencia del grupo de los homínidos (Tomada de Samonte y Eichler, 2002).

De acuerdo al histograma mostrado en la Figura 18, en el periodo comprendido entre los 13.39 y 15.62 ma sólo hay aparición de una retrosecuencia, lo cual puede deberse a que en efecto sólo se generó esa retrosecuencia en ese periodo histórico o que los otros retros que pudieron haberse generado simplemente se perdieron en algún momento. Se debe tener en cuenta que los retropseudogenes observados en la actualidad son el resultado no sólo de la retrotranscripción del mensajero y su reincorporación al genoma, si no también de una serie de factores que determinan que la retrosecuencias se hayan conservado hasta la actualidad, como son:

- a) Que la retrosecuencia se haya insertado en una región donde no cause alteraciones deletéreas.
- b) Que la retrosecuencia ya insertada se conserve y no sea alterada o fragmentada al grado de perder toda semejanza con la secuencia original.

- c) Que se dé en línea germinal.
- d) Que pase a las siguientes generaciones y se fije en la población.

En principio se podría pensar que los procesos de retrotranscripción no serían observados dadas las complicaciones a las que se enfrentan las retrosecuencias, pero la práctica nos dice que no es así y esto posiblemente se deba en buena medida a la baja cantidad de DNA codificante en el genoma y a la alta producción de retrosecuencias en momentos específicos.

Por otro lado al no estar sujetos a presiones de selección, los retroseudogenes tienden a divergir del gen original de varias maneras. Una de estas es la tendencia a disminuir su porcentaje de GC conforme transcurre el tiempo. Dado lo anterior se realizó una correlación entre el %GC y el tiempo de generación de cada secuencia. Como se observa en la Figura 20, existe una correlación negativa significativa entre el tiempo de aparición de las retrosecuencias y su %GC, es decir, al transcurrir un mayor tiempo de divergencia las secuencias van perdiendo GC.

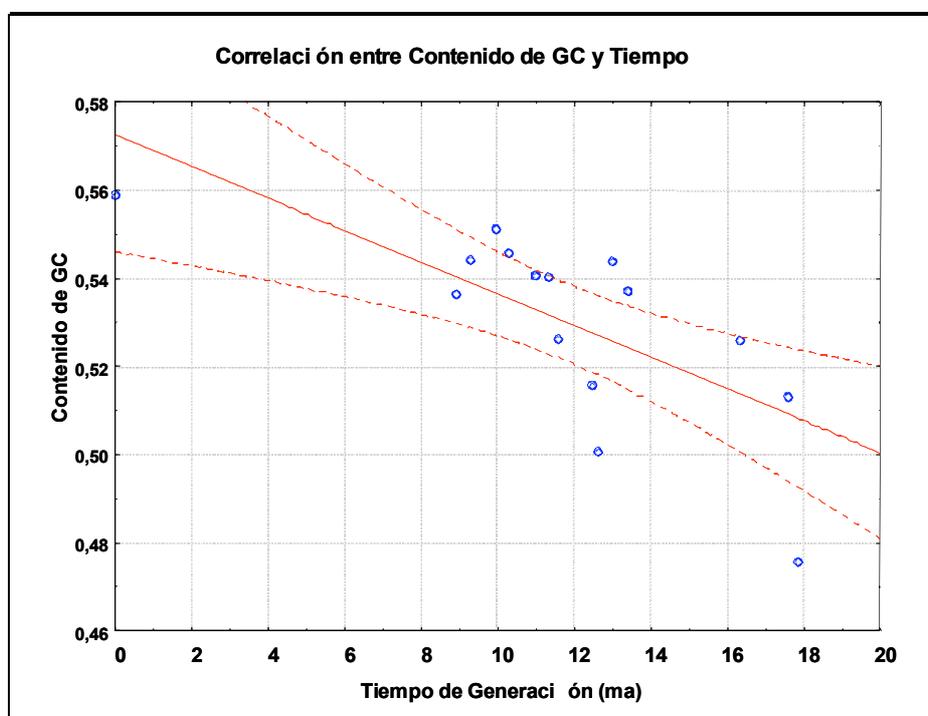


Figura 20. Correlación entre la proporción de GC y el tiempo de aparición de los retroseudogenes.

f. Determinación del potencial codificante de las retrosecuencias

Como se ha mencionado anteriormente, las retrosecuencias no son funcionales al momento de integrarse al genoma debido a que no contienen regiones promotoras ni reguladoras, pudiendo llegar a ser transcritas sólo si se insertan cerca de un promotor adecuado. Otro punto a considerar en la posible transcripción es la tasa de mutación de las secuencias, donde las retrosecuencias no sujetas a presiones de selección tendrán tasas de mutación mayores a su contraparte funcional, ganando así mutaciones que cambien el marco de lectura o generen codones de termino prematuros. Por tanto se parte de la premisa que las retrosecuencias analizadas no son funcionales a menos que aporten información que permita considerar lo contrario.

Para obtener los posibles marcos de lectura abiertos de las retrosecuencias se utilizó el programa Artemis. Las secuencias resultantes de la traducción en los tres marcos de lectura (sólo se tomaron los marcos que van de 5' a 3') para todos los retroseudogenes presentan una gran cantidad de codones de término, incluso las secuencias más recientes tienen codones de término prematuros, lo cual indica la ausencia de presión de selección por la acumulación de mutaciones en 1^a, 2^a y 3^a posición de manera indiscriminada, inserciones y deleciones. De esta manera según el programa Artemis, todas las retrocopias generaron posibles transcritos de tamaños muy inferiores a la secuencia original, que van del 28.15 al 78.15% de su tamaño (Figura 21).

La retrosecuencia denominada P11 es la única con un ORF de tamaño considerable –322aa–, pero solo representa el 78.15% del tamaño total de la secuencia original, por lo tanto si llegase a traducirse generaría un polipéptido corto no funcional, los otros polipéptidos presentaron tamaños que oscilan entre los 116 a 231 residuos de aa.

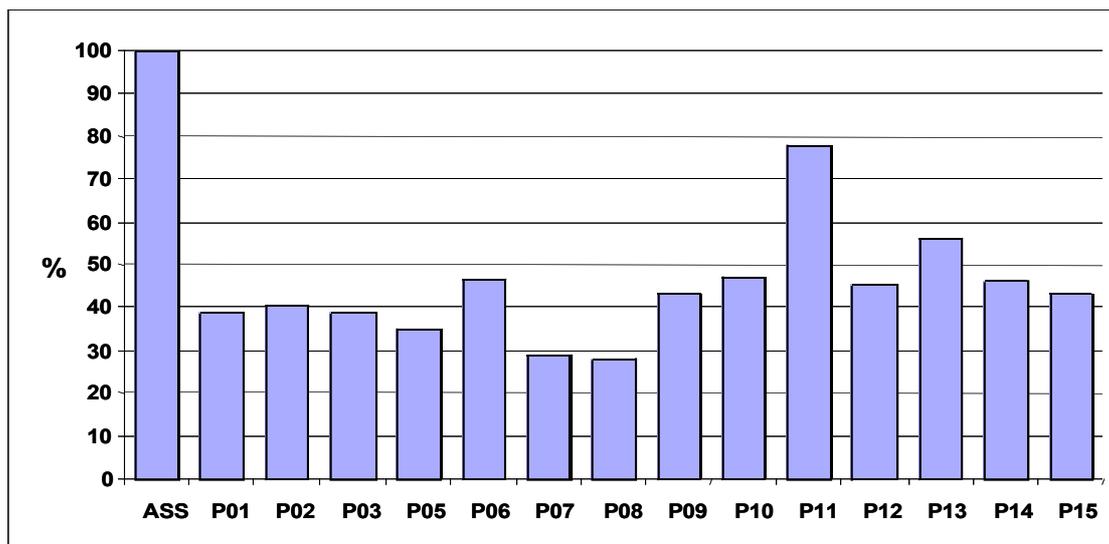


Figura 21. Proporción de tamaños de los polipéptidos predichos por Artemis en comparación al funcional.

Con base en lo anterior es poco probable que alguna de las secuencias se transcriba. Asimismo no hay indicios en las bases de datos de un transcrito parecido a ASS. Por tanto, todas ellas se pueden considerar como retropseudogenes.

Por otro lado, los retropseudogenes pueden ser considerados como fósiles moleculares, los cuales nos permiten hacer inferencias de los procesos históricos que han ocurrido desde su aparición, de esta manera se procedió a analizar cómo, debido a las mutaciones, las secuencias han cambiado de marco de lectura. Esto se llevo a cabo reconstruyendo los polipéptidos sin tener en cuenta los codones de término (Figura 22).

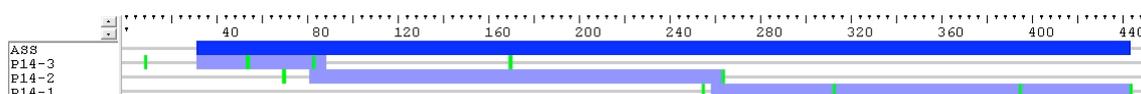


Figura 22. Reconstrucción de la secuencia polipeptídica original a partir de los tres marcos de lectura, las muescas señalan codones de término.

Las secuencias polipeptídicas derivadas son un mosaico de fragmentos que han cambiado una o varias ocasiones de marco de lectura como consecuencia de mutaciones, inserciones y/o deleciones (indels). A pesar de esto, aun conservan un alto grado de identidad con la secuencia original, teniendo porcentajes de identidad relativa de 73.37 a 89.68% con respecto al gen original, como resultado de su temprana aparición. Como cabría esperar según la teoría de reloj molecular, al comparar el tiempo de divergencia con el porcentaje de identidad relativa, como se observa en la Figura 23,

existe correlación negativa significativa entre estas dos variables, es decir, de manera general los retros más antiguos (P06, P02 y P15) presentan las menores identidades y aquellos más recientes (P09, P01 y P13) son los más parecidos a la secuencia original.

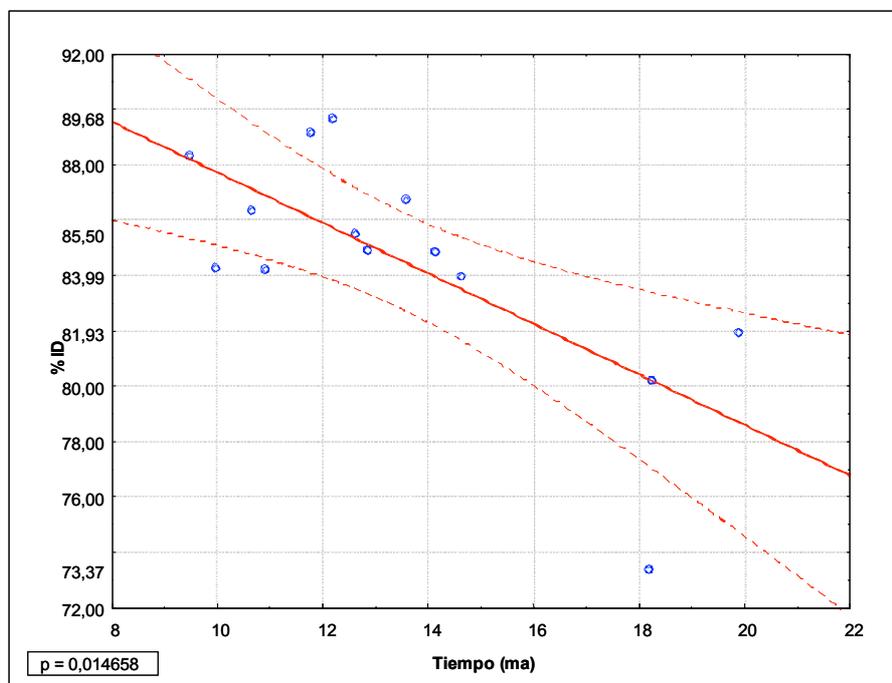


Figura 23. Correlación negativa entre el tiempo de divergencia y el % de identidad de los polipéptidos reconstruidos.

g. Análisis de polimorfismos en las secuencias

Según los resultados anteriores, como hipótesis de trabajo, se puede esperar que las diferencias entre el cDNA y los retros se deban a cambios en estos últimos. Sin embargo, se observó que en algunas posiciones quien mostraba el cambio era el cDNA siendo los retros parecidos entre sí (Tabla 3). Estas observaciones sugieren que el gen de ASS que dio origen a los retros sufrió cambios posteriores a la aparición de éstos, generando así la secuencia funcional que conocemos en la actualidad. Para tener una idea más amplia de los procesos ocurridos se procedió al análisis de los sitios y de ésta manera determinar donde se daban estos cambios y si dieron origen a diferencias en el polipéptido o no.

Sitio	9	56	155	203	545	587	1258	1345	1558
ASS Homo	C	A	G	T	T	G	C	G	A
ASS Pan	C	G	G	T	C	G	C	C	A
P01	T	G	C	C	C	A	T	C	-
P02	T	G	C	C	T	A	C	C	-
P03	T	G	C	C	C	A	T	C	-
P04	T	G	C	C	C	A	T	C	-
P05	T	G	C	C	C	A	T	C	-
P06	T	G	C	C	C	A	T	C	-
P07	T	G	C	C	C	A	T	C	-
P08	-	-	-	-	-	-	T	C	-
P09	T	G	C	C	C	A	T	C	-
P10	T	G	C	C	C	A	T	C	-
P11	T	G	G	C	C	A	T	C	-
P12	T	G	T	C	C	A	T	C	-
P13	T	G	C	C	C	A	T	C	-
P14	T	G	C	C	T	A	T	C	-
P15	-	-	-	-	-	-	T	A	-
Posición AA	5' UTR	5' UTR	3 ^a Ser	3 ^a Tyr	3 ^a Pro	3 ^a Glu	2 ^a Thr Ile	3' UTR	3' UTR

Tabla 3. Relación de los sitios de cambio entre los retros y la ASS de humano y chimpancé.

Se tuvo en cuenta que la argininosuccinato sintetasa es una enzima altamente conservada entre diversos grupos como: humanos, ratón, rata, bovinos e incluso levaduras (ver Tabla 1 en Introducción); esto significa que a lo largo de la evolución los cambios han estado altamente restringidos debido a que la actividad de esta enzima es vital para los organismos. En el humano las mutaciones en este gen producen alteraciones en el ciclo de la urea que pueden llevar a trastornos leves o irreversibles como retraso mental e incluso la muerte del individuo. A este tipo de trastornos asociados con la ASS se le conoce como citrulinemia (Kobayashi *et al*, 1990).

Del análisis realizado se deducen los siguientes puntos (ver Tabla 3):

- Los cambios se dan a lo largo de toda la secuencia, tanto en las regiones traducidas como en 5' y 3' UTR.
- Los cambios observados en la secuencia funcional de humano fueron posteriores a la aparición de los retros, es decir hace menos de 9 ma. Por tanto la secuencia que presentan los retros es la ancestral.
- La mayoría de ellos son cambios silenciosos, es decir ocurren en la tercera posición sin afectar con ello el aminoácido resultante.

- Los cambios presentes en la secuencia funcional no perjudicaron la funcionalidad de ASS, de lo contrario hubiesen sido seleccionados negativamente.
- Ninguno de estos cambios se correlaciona con los cambios que conducen a citrulinemia descritas por Kobayashi *et al* en 1990.
- La sustitución de T a C en el nucleótido 1258 en segunda posición de un codón que da origen al cambio de I a T en el aminoácido 386, es un cambio no conservativo. Pero al analizar otras especies se observa que la levadura también presenta T en dicha posición, por lo que, esta posición puede aceptar ambos aminoácidos sin que se vea perjudicada la funcionalidad de la enzima.
- Algunos de los retros sufrieron mutaciones en los mismos sitios, pero es poco probable que tengan una relación con los cambios en la secuencia funcional, siendo estos eventos independientes.

Para determinar si estos cambios habían ocurrido exclusivamente en humano o se dieron en otros homínidos se comparó la secuencia de ASS de humano contra la secuencia de ASS de chimpancé. Al realizar dicha comparación se observa que:

- Se presentaron cambios que afectaron tanto a la secuencia de chimpancé como a la de humano, lo cual indica que son cambios anteriores a su divergencia pero posteriores a la aparición del último retropseudogen (entre 6 a 9 ma).
- Existen también cambios que son humano específicos como son los de los sitios 56, 545 y 1345, posteriores a la divergencia humano-chimpancé.

No se han encontrado reportes comparativos de la actividad de ASS entre humano y chimpancé o con otros homínidos que permitan decir si estos cambios en el linaje del humano representa una diferencia en la actividad enzimática de la ASS, pero proveen una base para conocer la secuencia ancestral de la ASS y los cambios que se han dado desde la aparición del último retropseudogen en el linaje del humano.

h. Análisis de la vecindad metabólica de la ASS

Habiendo descrito y analizado los retropseudogenes asociados a la ASS, surgió la duda de cómo se comportarían las enzimas íntimamente asociadas a ésta en función a su número de retropseudogenes, en las diferentes rutas metabólicas en las que participa (ver introducción). Así comenzó la búsqueda de las retrosecuencias de las enzimas enumeradas a continuación:

- Carbamoil-fosfato sintetasa I (CPS I)
- Argininosuccinato liasa (ASL)
- Glicina amidinotransferasa (GATM)
- Arginasa I (tipo hígado)
- Arginasa II
- Ornitina carbamoiltransferasa (OTC)
- Óxido nítrico sintetasa Tipo I (Neuronal NOS, nNOS, NOS I)
- Óxido nítrico sintetasa Tipo II (Hepatocito NOS, HEP-NOS, NOS II)
- Óxido nítrico sintetasa Tipo III (Endotelial NOS, eNOS, NOS III)

Carbamoil-fosfato sintetasa (CPS I)

La CPS I, es la primera enzima en el ciclo de la urea, la cual forma carbamoil fosfato, a partir del amonio e iones bicarbonato con energía proveniente de dos moléculas de ATP. Esta reacción es esencialmente irreversible y es un paso limitante en el ciclo de la urea. Existe otra CPS la denominada CPS II, la cual es citoplasmática a diferencia de CPS I que esta en la mitocondria y es codificada por el gen CAD participando en la síntesis de pirimidinas (Martín-Hernández, 2005).

El gen de CPS I se localiza en el cromosoma 2 en la región 2q35, presenta un total de 38 exones que generan un transcrito de 5,743pb con un polipéptido resultante de 1,500 residuos de aminoácidos. Su número de acceso a las bases del ENSEMBL es ENSG00000021826.

El Blast-n mostró 43 hits todos ellos en la región donde esta localizado el gen del CPS I en el cromosoma 2. Al analizarlos se observó que había una concordancia de estos con los 38 exones del gen, por tanto se tomaran como un único hit no encontrando pseudogenes ni retrosecuencias (Figura 24).

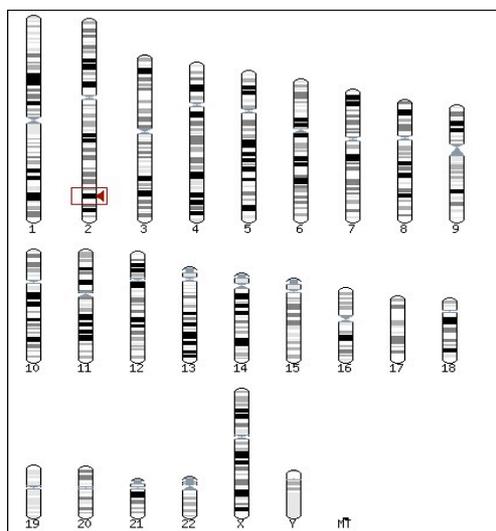


Figura 24. Blast para CPS I.

Argininosuccinato liasa (ASL)

Esta enzima está íntimamente relacionada con la ASS en las tres rutas donde participa, siendo la encargada de transformar el argininosuccinato en arginina y fumarato. Se trata de un gen de copia única, localizado en el brazo largo del cromosoma 7 en la región 7q11.21. Cuenta con 17 exones, un tamaño de transcrito de 1,509pb y genera un polipéptido de 464 residuos de aa. Su número de acceso en las bases del ENSEMBL es ENSG00000169910.

Al igual que en el caso de la ASS y las globinas, se mandó un Blast-n utilizando el cDNA como molde con diferentes valores para el parámetro “e”. Como resultado se obtuvieron dos hits uno de ellos corresponde al gen y otro hit en la región q11.23 del cromosoma 22 (ver Figura 25).

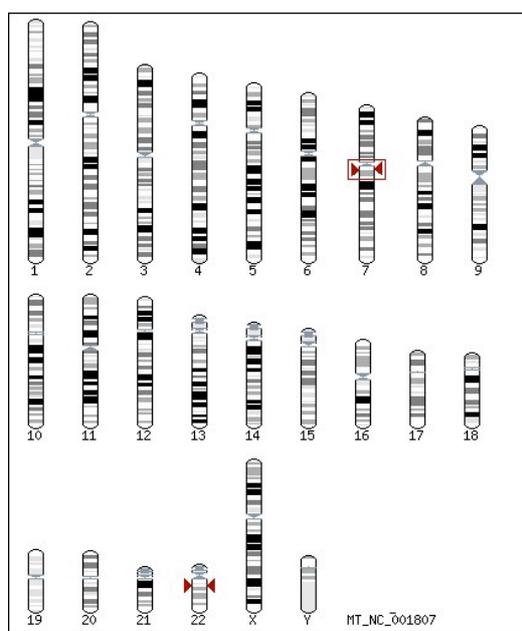


Figura 25. Resultante del Blast-n para ASL.

Al analizar el hit del cromosoma 22 para ver si se trataba de un retro o de un pseudogen se encontró que la secuencia recuperada presentaba fragmentos de diversos intrones, por lo cual este pseudogen no fue generado por un proceso de retrotranscripción, sino por un proceso de duplicación. Tampoco aparece en las bases de datos de Gerstein (www.pseudogen.org). Por tanto se puede decir que ASL no presenta retropseudogenes.

Glicina amidinotransferasa (GATM)

Esta enzima participa en la conversión de arginina a creatinina. Se encuentra en el cromosoma 15, en la región 15q21.1, su número de acceso es ENSG00000171766. El gen presenta, una sola copia con 9 exones, el tamaño del transcrito es de 2,346pb, dando lugar a un polipéptido de 423 residuos de aa.

El blast resultante solo reconoció al gen de la GATM (Figura 26) aunque los valores de “e” se modificaron. Por lo cual este enzima tampoco parece tener retrosecuencias relacionadas.

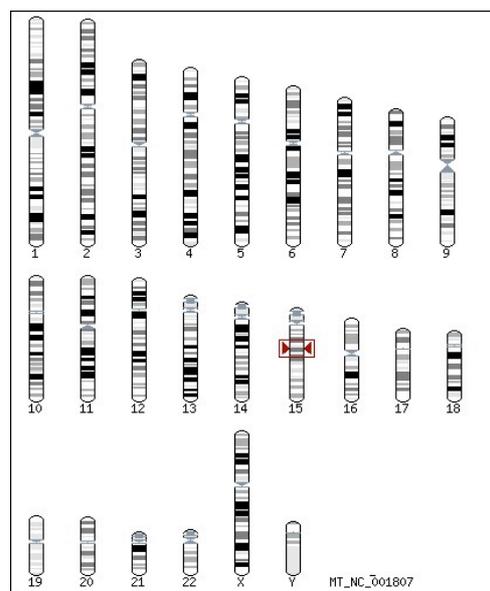


Figura 26. Blast-n de GATM.

Arginasa I (tipo hígado)

La arginasa I cataliza la hidrólisis de arginina a ornitina y urea. Se diferencia de la arginasa tipo II por su distribución en los tejidos, localización subcelular, reacción inmunológica y función fisiológica. Esta enzima es citosólica, se expresa principalmente en el hígado y es parte integral del ciclo de la urea. Esta localizada en el cromosoma 6, en la región 6q23. Presenta 8 exones con un transcrito de 1,471pb, generando un polipéptido de 330 residuos de aa. Su número de acceso es ENSG00000118520.

El blast arrojó dos hits, donde uno de ellos es la propia enzima y el otro es la arginasa tipo II presente en el cromosoma 14 (Figura 27).

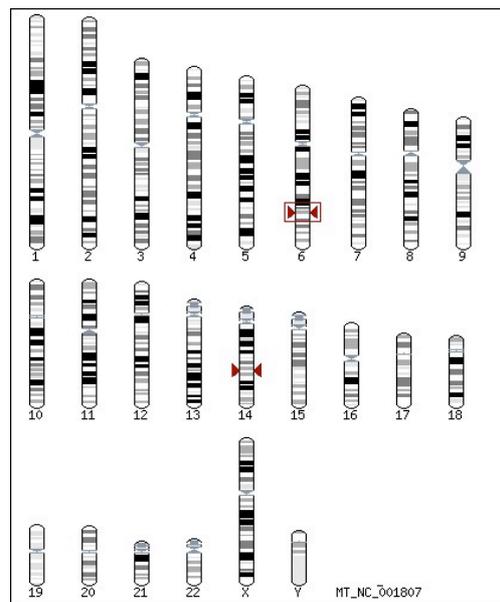


Figura 27. Blast resultante de Arginasa I.

Arginasa II

Esta enzima esta pobremente estudiada pero se cree que participa en el metabolismo del óxido nítrico y las poliaminas. Se localiza en la mitocondria de tejidos extra-hepáticos principalmente riñones. Su gen se encuentra en el cromosoma 14 en la región 14q21.1-q21.3. Cuenta con 8 exones, un tamaño de transcrito de 1,978 pb y un polipéptido de 354 residuos de aminoácidos. En el ENSEMBL presenta el número de acceso ENSG00000081181.

Del Blast se obtuvieron 7 hits (ver Figura 28). Donde los ubicados en los cromosomas 6 y 14 pertenecen a la arginasa I y II respectivamente. Los hits del cromosoma 3, 12 y los del 4 parecen ser un dominio de la arginasa II que se repite dentro de otras enzimas. Debido a que no es objetivo del presente trabajo determinar la estructura de los dominios de esta enzima así como tampoco ubicarlos dentro del genoma, no se continuó el análisis. Por ultimo, el segundo hit en el cromosoma 6 es una región pequeña de 35 nt y se trata de una región de baja complejidad integrada en su mayor medida por G y A que no parece ser perteneciente a una retrosecuencia.

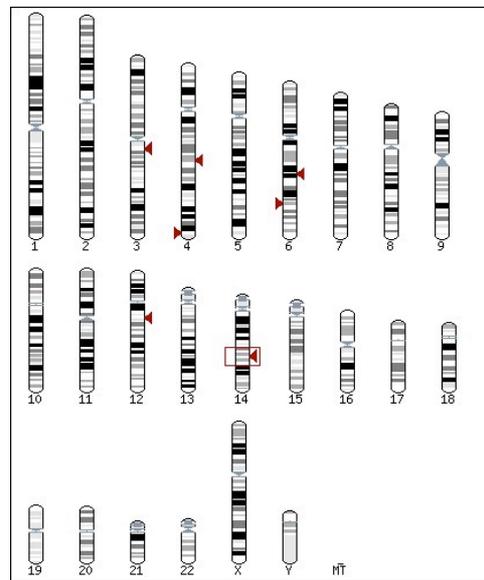


Figura 28. Patrón de hits encontrados para Arginasa II.

Ornitina carbamoiltransferasa (OTC)

La OTC cataliza la reacción de ornitina y carbamil fosfato para formar citrulina, restituyéndose así el ciclo de la urea. Esta enzima se encuentra ubicada en el brazo corto del cromosoma X, en la región Xp21.1. El gen esta conformado por 10 exones, con un transcrito de 1,931nt y un polipéptido de 354 residuos de aa. El número de acceso al gen es ENSG00000036473 para el ENSEMBL.

Al igual que en casos anteriores, el Blast localizó al gen de OTC y otros hits en el cromosoma 15 que pertenecen a secuencias pequeñas de no más de 25 nt (Figura 29). Por lo cual tampoco parece tener retrosecuencias asociadas.

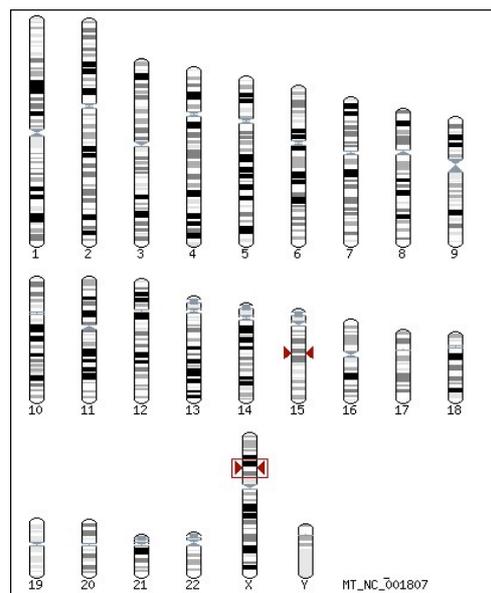


Figura 29. Hits encontrados para OTC.

Óxido nítrico sintetasa Tipo I (Neuronal NOS, nNOS, NOS I)

Las óxido nítrico sintetetas son una familia de enzimas que se encargan de producir óxido nítrico y citrulina a partir de arginina. Cada una de ellas puede presentar más de una isoforma activa. Debido a esta diversidad, se decidió analizar una isoforma de cada una de las tres variantes con el fin de evitar pasar por alto retros pertenecientes a una variante particular, ya que la bibliografía consultada muestra que en tejido gonadal tanto femenino como masculino pueden presentarse las tres variantes (Burnett et al, 1995 y Kim et al, 2005)

La nNOS, como su nombre lo indica, esta presente en el tejido neuronal, donde funciona como un señalizador de vida corta para el traspaso de impulsos nerviosos (Bredt *et al*, 1990). Se encuentra presente en el cromosoma 12, en la región 12q24.2 - 12q24.31. En las bases de datos se describen dos isoformas, la primera presenta 29 exones, un transcrito de 7,132pb y genera un polipéptido de 1,432 residuos, mientras la segunda isoforma presenta 28 exones con un transcrito de 8,646pb y un polipéptido de 1,436 residuos. Como se puede observar en la Figura 30, la NOS I presenta solo concordancia con ella misma así como con los miembros de su familia, presentes en los cromosomas 7 y 17. Aun variando los parámetros del Blast no se localizó ninguna retrocopia o pseudogen de esta.

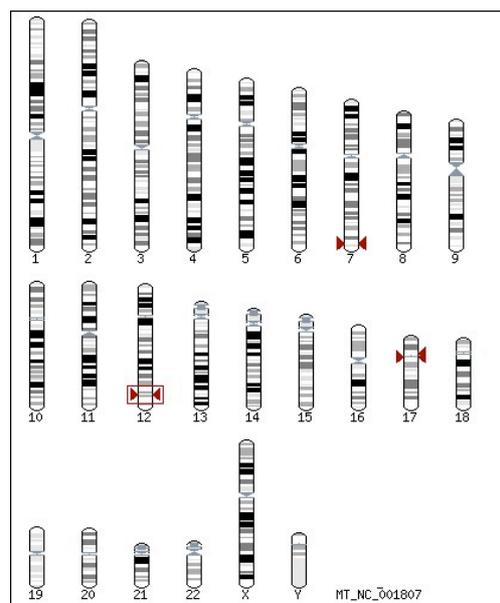


Figura 30. Blast-n para NOS I.

Óxido nítrico sintetasa Tipo II (Hepatocito o inducible NOS, iNOS, NOS II)

Esta enzima se expresa principalmente en el hígado aunque puede presentarse en otras células y tejidos en respuesta a endotoxinas y citocinas, a diferencia de NOS I debe ser inducida por una combinación de lipopolisacaridos y algunas citocinas. Esta presente en el cromosoma 17, región 17q11.2-q12 (Park *et al*, 1997). El gen se encuentra constituido por 27 exones, el tamaño del transcrito es de 4,203pb con un polipéptido de 1,153 residuos. Su número de acceso del gen es ENSG0000007171.

En la bibliografía se reporta que este gen presenta por lo menos un pseudogen no procesado cercano al gen original (Park *et al*, 1997), pero el análisis realizado muestra la presencia de tres pseudogenes no procesados cercanos al gen original, lo cual se vio confirmado al revisar la base de datos del NCBI para dicho gen. Como se muestra en la Figura 31, se da un reconocimiento de los otros miembros de su familia pero existen hits en los cromosomas 2, 14 y 22 que no son retrocopias y podrían corresponder al dominio de flavodoxina_1 presente en la proteína. Por tanto esta variante tampoco presenta retroseudogenes.

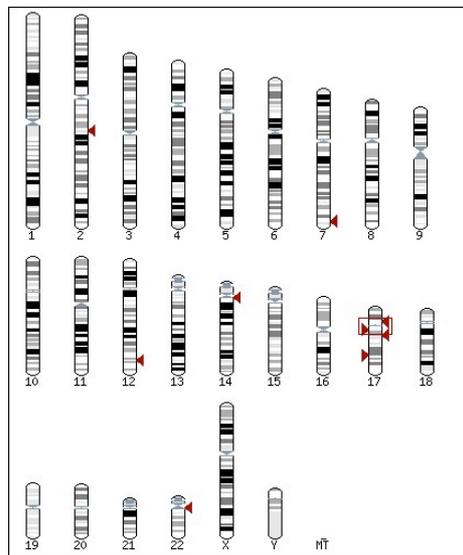


Figura 31. Hits resultantes para NOS II.

Óxido nítrico sintetasa Tipo III (Endotelial NOS, eNOS, NOS III)

Al igual que nNOS, la eNOS es una enzima constitutiva que requiere para su activación calcio/calmodulina aunque se ha observado que puede incrementarse su actividad y cantidad sugiriendo con ello que es capaz de inducirse (Texeira, 2003). Se expresa en las células del tejido endotelial, cabe destacar que cada una de las tres variantes de NOS se encuentra localizada en distintas regiones dentro de la célula, lo cual puede estar relacionado con los diferentes mecanismos de regulación. La eNOS se encuentra localizada en regiones específicas de la membrana plasmática (Wu y Morris, 1998).

Se localiza en el cromosoma 7 en la región 7q36, está constituido por 27 exones, presenta un transcrito de 4,323pb y un polipéptido de 1,203 residuos. Su número de acceso en la base del ENSEMBL es ENSG00000164867.

Para llevar a cabo el estudio de esta variante se tuvo que hacer un ajuste eliminando los primeros 1000nt del cDNA molde debido a que esta enzima presenta una región rica en prolina que interfería en el estudio por ser muy abundante en el genoma. El resultante fue la localización del gen y sus variantes, sin la aparición de retros o pseudogenes no procesados (Figura 32)

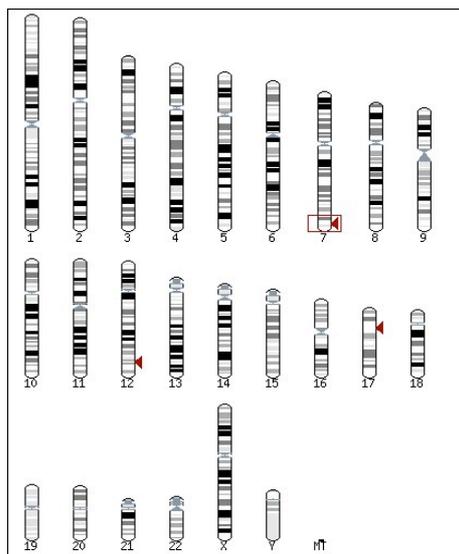


Figura 32. Patrón de hits para NOS III.

La siguiente tabla resume los resultados de los genes analizados.

Crom	No. de Enzima	Nombre	No de copias encontradas
9	6.3.4.5	Argininosuccinato sintetasa (ASS)	14 retropseudogenes
7	4.3.2.1	Argininosuccinato liase (ASL)	1 pseudogen (Cromosoma 22)
15	2.1.4.1	Glicina amidinotransferasa (GATM)	0
6	3.5.3.1	Arginasa I (tipo hígado)	0
14	3.5.3.1	Arginasa II	0
X	2.1.3.3	Ornitina carbamoiltransferasa (OTC)	0
12	1.14.13.39	Óxido nítrico sintetasa Tipo I (NOS I, Neuronal NOS, nNOS)	0
17	1.14.13.39	Óxido nítrico sintetasa Tipo II (NOS II, Hepatocito ó inducible NOS, iNOS)	3 pseudogenes (Cromosoma 17)
7	1.14.13.39	Óxido nítrico sintetasa Tipo III (NOS III, Endotelial NOS, eNOS)	0
2	6.3.4.16	Carbamoil-fosfato sintetasa (CPS I)	0

Tabla 4. Resumen de la relación de retropseudogenes y pseudogenes asociados con las enzimas de la vecindad metabólica de ASS.

En la Tabla 4 se observa, de manera generalizada, que ninguna de las enzimas o sus variantes asociadas a la ASS presentan retropseudogenes. Como ya se había mencionado la premisa principal para la formación de retropseudogenes es que el nivel de expresión del gen se eleve en células germinales. ¿Pero cómo se puede explicar que solo la ASS haya aumentado su nivel de expresión hace aproximadamente 18 hasta 8 ma y las otras enzimas altamente relacionadas con ella no?

En la actualidad se sabe que la ASS puede localizarse en diferentes regiones de la célula dependiendo del ciclo en el que participe. Como parte del el ciclo de la urea la ASS se encuentra en la mitocondria mientras que para la síntesis de arginina y NO se encuentra en diversas regiones del citosol (Wu y Morris, 1998). Además de sus diversas localizaciones espaciales también presenta una alta diversidad de relaciones enzimáticas. En gónadas de algunos mamíferos (ratas, ratones y cerdos) se ha observado que la ASS esta actuando en el ciclo productor de NO, el cual se ha denominado ciclo citrulina/arginina o citrulina/NO, en dicho ciclo se encuentra estrechamente relacionada con las tres variantes de NOS (neuronal, epitelial e inducible) pues las tres se presentan en dichos órganos (Lissbrant *et al*, 1997, Kim *et al*, 2005 y Burnett *et al*, 1995). Sin embargo, en los tres casos la interacción no se da de la misma manera, observándose que, en la síntesis de NO –con la ayuda de ASS e iNOS– estas dos enzimas son co-inducidas (Morris, 2002).

Dado lo anterior se puede suponer que la aparición de la ASS en el tejido gonadal es debido a la producción de NO, el cual se utiliza en el proceso de desarrollo folicular y/o en los procesos ovulatorios (Kim *et al*, 2005), en los testículos está involucrado en el control de la vasculatura testicular (Lissbrant *et al*, 1997) y en los movimientos contráctiles, hemodinámica y procesos secretorios en el tracto reproductivo masculino, lo cual implica su participación en la regulación de los mecanismos requeridos para la función reproductiva masculina (Burnett *et al*, 1995).

Otro punto a considerar es que la ASS se trata de una enzima tasa limitante (rate-limiting) debido a que la formación del argininosuccinato es termodinámicamente desfavorable y lenta, necesitando ATP. La reacción de unión de la citrulina y aspartato para formar argininosuccinato se lleva a cabo en dos pasos con el gasto de ATP. En primer lugar se da la activación de la citrulina al unirse al ATP formando citrulin-adenilato activado, con desprendimiento de pirofosfato, la velocidad de este paso es de 6.2 s^{-1} . Posteriormente se da el ataque nucleofílico del grupo amino del aspartato contra el citrulin-adenilato, generándose así el argininosuccinato y AMP residual; la velocidad de este paso (0.9 s^{-1}) es siete veces menor que la del paso anterior, lo que sugiere que el segundo paso es el limitante. A pesar de esto, la formación del citrulin-adenilato tiene que darse en presencia del aspartato de lo contrario la velocidad de reacción sería 600 veces menor que en presencia de este (0.01 s^{-1}) (Lemke & Howell, 2002 y Ghose & Raushel, 1985).

Conociendo las características anteriormente descritas de la ASS, se puede inferir que entre hace 18 y 8 ma aumento drásticamente el requerimiento de NO en gónadas lo que llevó a un aumento en la transcripción de ASS debido a que los principales mecanismos de regulación de ASS se encuentran a dicho nivel, no siendo el mismo caso de ASL y NOS. Su aumento tuvo que ser aun más pronunciado debido a que se trata de una enzima tasa limitante lo que quiere decir que este solo paso enzimático aletarga el resto del ciclo, por tanto, se debe aumentar la cantidad de enzima para compensar su lentitud, sin que necesariamente haya aumentado el nivel de expresión de sus enzimas compañeras en el ciclo citrulina/NO (ASL y NOS), pues como ya se ha mencionado, pueden o no estar correguladas. Cabe también la posibilidad que ASL y NOS también hayan aumentado su transcripción pero de manera tan baja que no dio la posibilidad de que el mRNA se uniera a la reversotranscriptasa.

i. Análisis de la vecindad evolutiva de la ASS humana

Análisis de los retropseudogenes asociados a ASS en el genoma del chimpancé

Teniendo a disposición en las bases de datos el borrador del genoma del chimpancé, se realizó la búsqueda de los retropseudogenes asociados a ASS en el genoma de dicho organismo.

Chimpancé y humano compartieron su último ancestro común entre 4.6 y 6.2ma, lo cual es muy reciente en términos evolutivos, por lo cual ambos genomas son altamente parecidos (Ebersberger *et al*, 2002). La sinténia de ambos es muy elevada, teniendo una relación 1:1 de los cromosomas –a excepción de los cromosomas 12 y 13 de chimpancé que en el humano se fusionaron para dar origen al cromosoma 2– (Feuk *et al*, 2005).

La ASS de chimpancé se encuentra, como se esperaba, en la región sinténica con el humano –el antes llamado cromosoma 11 ahora cromosoma 9–. Las bases de datos aun no tienen bien caracterizado este gen, al cual le asignan entre 10 y 11 exones con un tamaño de transcrito de 853nt y un polipéptido de 258 residuos de aa (Figura 33). Estos datos son incongruentes teniendo en cuenta que los polipéptidos de ASS de vaca, rata y ratón al igual que el del humano tienen 412 residuos de aminoácidos, por lo que se esperaba que chimpancé tuviera la misma longitud. Por otro lado, el cDNA de las bases de datos es altamente diferente del cDNA propuesto para humano y si se utiliza como molde para la predicción de un polipéptido genera una secuencia trunca por la aparición prematura de codones de término. Por lo tanto, se procedió a obtener la secuencia genómica de ASS de las bases de datos y hacer el empalme de exones de manera manual utilizando como molde el cDNA humano.

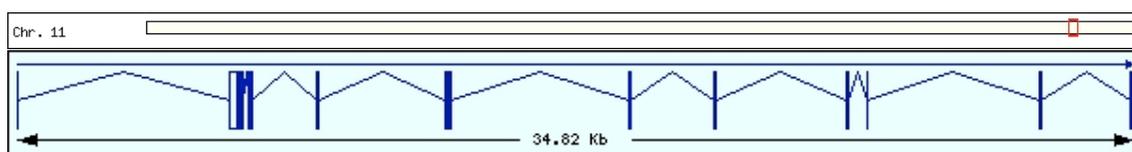


Figura 33. Patrón de exones predicho por ENSEMBL para ASS en chimpancé y su localización en el cromosoma.

Se generó un cDNA con una identidad de 99.05% en comparación al humano (Para ver la secuencia completa del cDNA ir al Anexo II). Dicho cDNA proviene de un gen con 16 exones con un tamaño de transcrito de 1598 nt contra 1592 nt de humano, esta mínima diferencia es debida a indels en las regiones 5' y 3' UTR. Al predecir el ORF de mayor tamaño y el polipéptido resultante a partir del cDNA modificado, se genera un polipéptido de 412 residuos aunque este tiene regiones incompletas pues el gen en las bases de datos aun no está bien anotado. El porcentaje de identidad de dicho polipéptido es de 99.01% comparado con el humano (Para ver la secuencia completa del polipéptido ir al Anexo III). Los dos cambios detectables en la secuencia de aminoácidos son:

Posición	66	138
Humano	R	Q
Chimpancé	K	H

Tabla 5. Cambios de aa entre humano y chimpancé.

Existen otros cambios pero estos no pueden ser concluyentes debido a la falta de información de la secuencia.

Habiendo caracterizado el cDNA de chimpancé y su polipéptido, se hizo la búsqueda de las retrosecuencias asociadas a este, obteniéndose un total de 63 hits. Al descartarse las secuencias cortas y el gen funcional resultaron 15 hits que, como se puede observar en la Figura 34, al alinear los cromosomas homólogos concuerdan perfectamente con los hits encontrados en humano. Únicamente el hit asociado al cromosoma humano 12 no tiene contraparte en el genoma del chimpancé, pero al obtener y depurar las secuencias se encontró que una secuencia denominada 5_random se trataba del homólogo al retropseudogen P14 de humano.

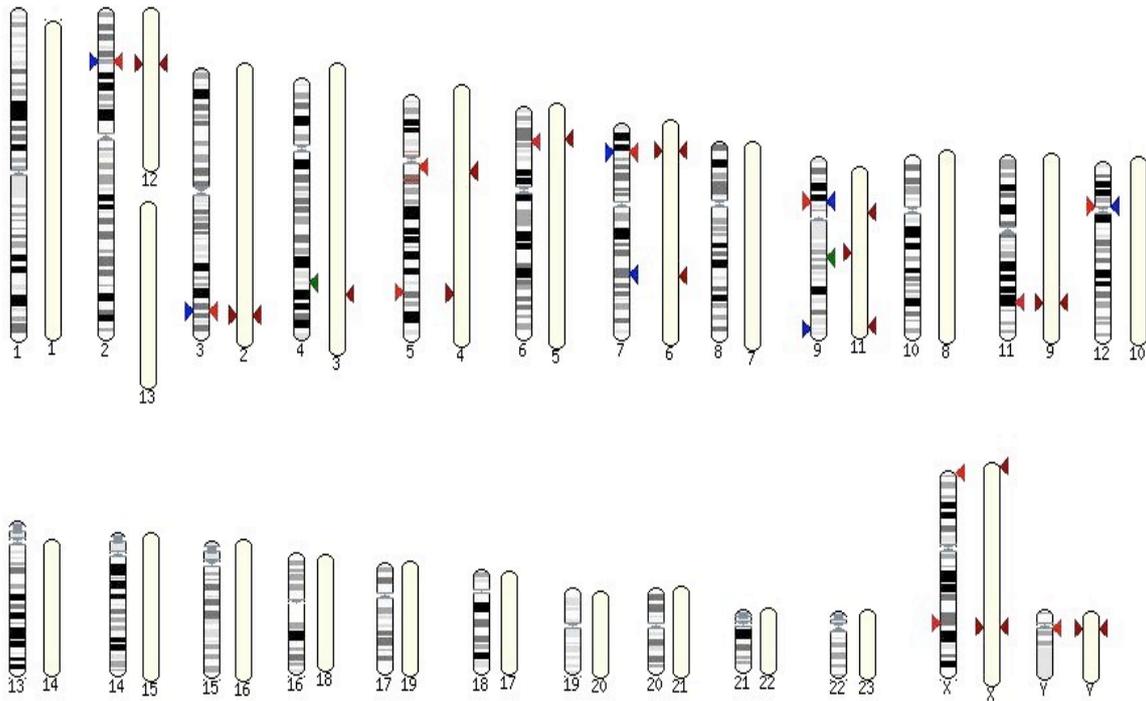


Figura 34. Comparación de los hits encontrados en humano (izquierda) y chimpancé (derecha). Se observa una concordancia 1:1 entre ellos.

Al analizar las 15 secuencias se observó que corresponden 1:1 a las secuencias encontradas en humano. Este dato concuerda con los datos de datación de las secuencias, donde todas ellas aparecieron entre 8.92 y 17.85 ma, es decir, el último retroseudogen apareció aproximadamente 2 millones de años antes de la divergencia humano-chimpancé, lo cual conlleva a encontrar los mismos retos en ambos genomas. El hecho de que existan regiones del genoma del chimpancé sin secuenciar limita el alcance de las secuencias comparadas.

Se puede esperar que las historias de aparición de los retroseudogenes de chimpancé ortólogos a los humanos sean muy parecidas, donde, una de las retrocopia encontradas en el cromosoma X sea en verdad un evento de duplicación y no un evento de retrotranscripción. Cabría esperar también, que los procesos evolutivos posteriores a la aparición de los linajes de humano y chimpancé fueran parecidos aunque esto no puede darse por hecho pues ambos genomas tienen tasas de cambio diferentes, por lo cual se tendría que hacer un estudio más detallado de los procesos que han sufrido los retroseudogenes de ASS en el linaje del chimpancé.

Otro punto importante son los cambios observados en la secuencia de nucleótidos de humano y chimpancé, teniendo como modelo de la secuencia ancestral a los retros. Como ya se había mencionado (Tabla 3) existen cambios posteriores a la aparición del último retro pero anteriores a la divergencia de ambos grupos, y otros que al parecer son humano específicos. También los chimpancés presentan cambios específicos en comparación a la secuencia ancestral y al genoma humano, de hecho, los aminoácidos diferentes entre humano y chimpancé (Tabla 5) son cambios que ocurrieron en el linaje de este último, siendo el cambio de R a K producto de un cambio en segunda posición, mientras el de Q a H fue en tercera.

Ambos cambios según las tablas de M. O. Dayhoff, son de tipo conservativo, lo cual muestra que ambos genomas están en constante cambio, pero dado que la ASS es una enzima vital, los cambios tienen que ser restringidos. El cambio de R por K en el aminoácido 66 de chimpancé es un cambio reciente pero al compararlo con otras especies incluyendo *E. coli* se observa que la lisina es un aminoácido altamente conservado, el cual forma parte de la alfa hélice 3 que sirve de enlace entre el llamado dominio de unión al nucleótido y el dominio de sintetasa (Lemke *et al*, 2001). Pero en algún momento en el linaje de los homínidos éste cambió por arginina restableciéndose en el linaje del chimpancé. Por su parte el cambio de Q a H parece ser un cambio novedoso ya que la histidina está altamente conservada entre los mamíferos como rata, ratón, humano y bovino. Éste residuo forma parte de una región entre la alfa hélice 5 y la hoja beta 6 en el dominio de unión al nucleótido (Lemke *et al*, 2001).

Como se pudo observar, el análisis de los retropseudogenes en ambos genomas permite por un lado saber si la calibración del reloj molecular es correcta pues se conoce por registros fósiles y moleculares los tiempos aproximados de divergencia entre organismos. Si este tiempo es menor a la aparición del último retro entonces todos los retros deben encontrarse en ambos genomas, a menos que alguno de ellos se haya perdido por procesos posteriores de recombinación o reacomodo del genoma. Por otro lado permite usar a los propios retropseudogenes como un registro fósil de la secuencia que dio origen a las ASS presentes en los dos linajes, así como a partir de estos conocer los cambios que han ocurrido en ambos linajes y saber concretamente si el cambio se dio en uno u otro linaje.

Análisis de los retropseudogenes asociados a ASS en el genoma del ratón

Otro organismo que se puede considerar de la vecindad evolutiva del humano es el ratón. El conocer su número de retrosecuencias asociadas a la enzima ASS nos puede decir mucho acerca de los procesos que se han dado en otros linajes diferentes al de los homínidos y si entre ellos se observan similitudes en la expresión a lo largo del tiempo de la misma enzima. Así, se buscaron las retrosecuencias de ASS en este organismo que además ha sido ampliamente utilizado como modelo evolutivo.

En las bases de datos del ENSEMBL para genoma de ratón (*Mus musculus*) el gen de la ASS se encuentra con el número ENSMUSG00000026844. Este gen está presente en el cromosoma 2 en la región 2B;220.0cM, cuenta con 16 exones, un tamaño de transcrito de 1631pb y un polipéptido de 412 residuos (Figura 35)

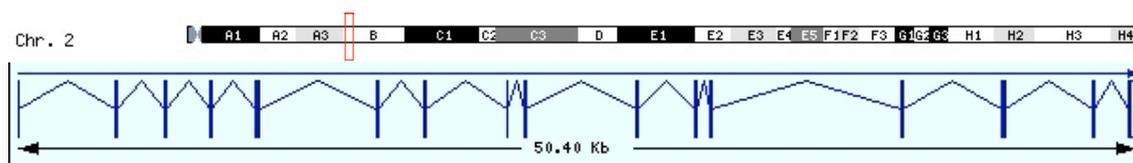


Figura 35. Localización de ASS en el cromosoma 2 de ratón y su patrón de exones.

Humano y ratón divergieron hace aproximadamente 85ma (Bromham *et al*, 1999), a pesar de esto, las secuencias de ASS se encuentran altamente conservadas. Al comparar los cDNA se obtiene un 86.25% de identidad, mientras que los polipéptidos presentan un 96.8%, lo cual refleja la importancia de esta enzima en los procesos de excreción de desechos nitrogenados, formación de arginina y NO.

La búsqueda de retrosecuencias arrojó 20 hits, de los cuales 16 de ellos correspondían al gen original (localizado en el cromosoma 2) y los cuatro restantes pertenecían a una misma secuencia en el cromosoma 10 (Figura 36). Al empalmar los hits presentes en el cromosoma 10 se obtuvo una secuencia de 1594 nt, al ser comparada contra el cDNA original se pudo observar que existía un gran parecido con éste con solo 6 cambios de nucleótidos, siendo el porcentaje de identidad entre ellas de 99.37%, aunque carecía de 32 pb en la región inicial 5', posiblemente debido a la caída de la reverso transcriptasa. Se pudo destacar la total ausencia de intrones o inserciones que pudieran ser vestigios de estos y la presencia de vestigios de una cola de poli A. Con

estos datos se puede asegurar que se trata de una retrosecuencia derivada del gen de ASS en ratón.

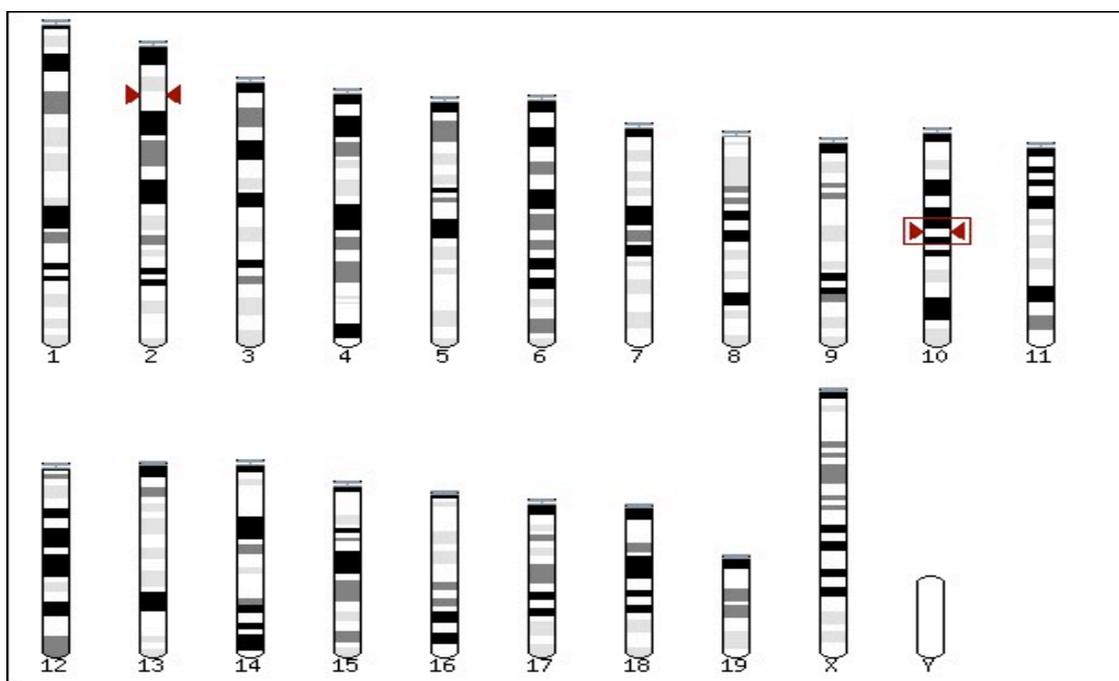


Figura 36. Hits encontrados para ASS en el genoma del ratón donde el hit del cromosoma 2 es el gen y el del 10 es la retrosecuencia.

Siguiendo el análisis de la retrosecuencia se hizo la traducción en los tres marcos de lectura para determinar su posible viabilidad, obteniéndose un polipéptido de 412 residuos de aminoácidos, es decir del mismo tamaño que el polipéptido funcional, el cual presentaba 99.51% de identidad con respecto a éste último. Al estudiar los cambios en la secuencia se pudo identificar solo dos cambios de aminoácidos: T por S en el aa 174 y S por T en el aa 219, ambos cambios, según las tablas de M. O. Dayhoff son conservativos (para ver el polipéptido predicho para la retrosecuencia ir al Anexo IV). Cabe destacar que como ya se mencionó, la secuencia de nucleótidos presentaba 6 cambios, lo cual indica que solo dos de estos cambios dieron origen a un cambio de aminoácido mientras los otros 4 fueron silenciosos, estas peculiaridades de la retrosecuencia con respecto a la secuencia original parecen ser debidas a presiones de selección y no a mutaciones al azar, por lo que dicha secuencia podría ser una retrocopia funcional del gen de la ASS. Sin embargo, el reducido número de cambios es evidencia de que esta retrosecuencia es muy joven y aun no ha acumulado mutaciones, inserciones y/o deleciones que indiquen claramente si es un gen procesado un retropseudogen.

Humano y ratón han tenido historias evolutivas diferentes a partir de su divergencia, lo cual se ve reflejado en la historia de la ASS en cada linaje. En humano esta enzima tuvo un periodo de aproximadamente 10 ma –hace 8 a 18 ma– donde su nivel de expresión se incrementó permitiendo de esta manera que se generaran retrosecuencias, en cambio en el linaje del ratón los niveles de expresión de la ASS en gónadas se mantuvieron constantes a lo largo del tiempo, únicamente con una variación que permitió la generación de una retrosecuencia que muestra alguna probabilidad de ser funcional, a diferencia de las retrosecuencias de humano que en su totalidad se tratan de retropseudogenes.

Conclusiones y perspectivas

La visión de los retropseudogenes como fósiles moleculares no es nueva, se les comenzó a atribuir dicho nombre por tratarse de remanentes de procesos que ocurrieron en el pasado, pero hasta últimas fechas se ha ido descubriendo el potencial de éstos en los estudios de evolución y dinámica de genes y genomas. Al mismo tiempo de ser modelos idóneos para el estudio de la dinámica de secuencias dirigidas por evolución neutral (Devor *et al*, 2003), los retropseudogenes permiten observar cambios en los niveles de expresión de genes a lo largo de su historia, como lo constatan trabajos anteriores, actuando como “paleomicroarreglos” que permiten saber en qué momentos históricos se dieron aumentos drásticos en los niveles de expresión de un gen como resultado de cambios externos (Barberán-Soler, 2002, Alcaraz-Peraza, 2005 y Vilchis-Peluyera, 2006).

Además de obtener información sobre los cambios de niveles de expresión en genes de proteínas específicas, en el presente trabajo se pudo constatar el uso de los retropseudogenes como “moldes” para visualizar dos aspectos importantes de la naturaleza de los genes implicados:

Por un lado el uso de la totalidad de retropseudogenes derivados de un gen particular permiten la reconstrucción de la variante de mRNA que les dio origen. Este aspecto es muy importante cuando se trabaja con genes que pueden presentar splicing alternativo con es el caso del gen de la argininosuccinato sintetasa, donde no se conocen hasta el momento todas las isoformas resultantes posibles. La reconstrucción del patrón de splicing y del cDNA resultante puede llevar a tres líneas de pensamiento:

- 1) El cDNA resultante es el producto de la expresión de una isoforma presente únicamente en gónadas.
- 2) El tipo de mRNA que les dio origen fue el que se expresaba en el linaje que después divergiría a humano y chimpancé hace entre 18 a 8 ma.

- 3) Dicha isoforma es la expresada únicamente en la ruta de síntesis del NO, ya que no hay ciclo de la urea en gónadas, lo que implicaría que el tipo de splicing tiene repercusiones regulatorias.

La limitante del estudio *in silico* es no poder determinar con certeza cuál o cuáles de los panoramas planteados son correctos, para esto se tendría que recurrir a estudios *in vivo* y/o *in vitro*.

El segundo aspecto para el cual los retropseudogenes pueden ser utilizados como moldes, es en la reconstrucción de la secuencia ancestral extinta que los originó. Conociendo las discrepancias entre los retropseudogenes y la secuencia funcional se puede determinar cuáles cambios se han dado en los retros como producto de mutaciones espontáneas y aquellos que se originaron en la secuencia funcional actual, pudiéndose reconstruir de esta manera la secuencia del gen ancestral que dio origen a los retros. Con base en esta reconstrucción se logra conocer cómo ha divergido la secuencia funcional actual a partir de esta secuencia ancestral. Además, si existió divergencia de linajes posteriores a la aparición de los retros se puede realizar el comparativo de los cambios que comparten los linajes y de los cambios linaje específicos o la aparición de homoplasias.

Inesperadamente, el estudio de los retropseudogenes y las copias originadas de éstos permitieron apreciar la dinámica de interacción entre cromosomas, pudiendo reconocer sitios de recombinación no homóloga entre los cromosomas sexuales del linaje de humano y chimpancé que sucedieron hace aproximadamente 14 ma.

Es necesario señalar que las enzimas no se encuentran actuando de manera independiente en el metabolismo, éstas se encuentran interrelacionadas estrechamente con otras, por lo cual el estudiar los retropseudogenes asociados a los genes de las enzimas presentes en una o más rutas metabólicas puede arrojar información sobre la manera de afrontar los cambios medioambientales por parte de una ruta metabólica. Este análisis llevó a descubrir que al presentarse una modificación en los requerimientos de una ruta, no todas las enzimas se comportarán de la misma manera a pesar de estar relacionadas, siendo sólo algunas –o en el presente caso sólo una– la que variará sus niveles de expresión para poder afrontar requerimientos metabólicos más pronunciados.

Esta enzima se puede denominar como una enzima nodo pues es la que aumentará sus niveles de expresión para activar de esta manera el ciclo y poder hacer frente a los requerimientos del momento.

Este tipo de estudios da un panorama novedoso de los procesos históricos que una ruta metabólica ha sufrido en función de los cambios en los requerimientos metabólicos, como producto de posibles cambios externos y podría aportar más información sobre la regulación de las rutas metabólicas donde sea posible emplearlo.

Al igual que los estudios paleontológicos los estudios con retroseudogenes se basan en toda la información previa disponible, es así que el análisis de la vecindad evolutiva de genes con presencia de retroseudogenes aporta información que permite por una parte corroborar la datación de las retrocopias al comparar la datación de estas contra los tiempos de divergencia de los organismos estudiados, como fue el caso de humano y chimpancé que al compartir todos los retroseudogenes de ASS corroboró la datación que ubicaba al más reciente aproximadamente 2 ma antes de la divergencia de los dos linajes. Pero a su vez la datación de las retrocopias podría proporcionar indicios de los tiempos de divergencia de linajes cuyas dataciones no sean del todo claras.

Por otro lado su estudio en linajes que divergieron en tiempos antiguos, como humano y ratón, muestra como en linajes independientes la evolución de los genes y los requerimientos de éstos a lo largo de la historia también son diferentes, producto de un desarrollo histórico divergente con presiones y cambios externos propios de cada linaje.

Hasta el momento únicamente se ha abordado el estudio de retroseudogenes, pero también existe la posibilidad de que este tipo de estudios descubran retrocopias funcionales, ya que se sabe que la retrotransposición de secuencias derivadas de genes funcionales es uno de los mecanismos por los cuales se pueden generar nuevos genes, en el caso de que la retrosecuencia resultante se reinserte en una región con un promotor propio o bajo la influencia de un *enhancer*. El encontrar retrocopias con potencial codificante mediante el análisis *in silico* podría ser el primer paso para descubrir genes codificantes que no han sido descritos en las bases de datos y poder analizar de cerca a estos genes “de reciente creación” y cómo pueden llegar a divergir del gen del cual se originaron.

De manera general el estudio de los retropseudogenes está comenzando a cambiar, de estudios meramente descriptivos a estudios donde éstos son utilizados como herramientas moleculares que permiten conocer más sobre la dinámica e historia de genes, genomas e incluso sobre rutas metabólicas específicas. De esta manera, el continuar el estudio de estos elementos moleculares, con la ayuda de la creciente información sobre genomas de primates, podría ir generando más información novedosa sobre la evolución de este linaje y los cambios que ha sufrido cada una de sus ramas filogenéticas.

ANEXOS

ANEXO I. Secuencia de cDNA modificada de humano

>cDNA ASS

1 CTGCTCTGCCGCTGCCACCGCTGCCCCGAGCCCGAGTGGTTCACT 45
 46 GCACTGTGAAAACAGATTCCAGACGCCGGAACTCACGCCTCCAA 90
 91 TCCCAGACGCTATGTCCAGCAAAGGCTCCGTGGTTCCTGGCCTACA 135
 136 GTGGCGGCCTGGACACCTCGTGCACTCCTCGTGTGGCTGAAGGAAC 180
 181 AAGGCTATGACGTCATTGCCTATCTGGCCAACATTGGCCAGAAGG 225
 226 AAGACTTCGAGGAAGCCAGGAAGAAGGCACTGAAGCTTGGGGCCA 270
 271 AAAAGGTGTTTCATTGAGGATGTCAGCAGGGAGTTTGTGGAGGAGT 315
 316 TCATCTGGCCGGCCATCCAGTCCAGCGCACTGTATGAGGACCGCT 360
 361 ACCTCCTGGGCACCTCTCTTGCCAGGCCCTGCATCGCCCGCAAAC 405
 406 AAGTGAAAATCGCCAGCGGGAGGGGGCCAAGTATGTGTCCCACG 450
 451 GCGCCACAGGAAAGGGGAACGATCAGGTCCGGTTTGAGCTCAGCT 495
 496 GCTACTCACTGGCCCCCAGATAAAGGTCATTGCTCCCTGGAGGA 540
 541 TGCCTGAATTCTACAACCGGTTCAAGGGCCGCAATGACCTGATGG 585
 586 AGTACGCAAAGCAACACGGGATTCCCATCCCGGTCCTCCCAAGA 630
 631 ACCCGTGGAGCATGGATGAGAACCTCATGCACATCAGCTACGAGG 675
 676 CTGGAATCCTGGAGAACCCCAAGAACCAAGCGCCTCCAGGTCTCT 720
 721 ACACGAAGACCCAGGACCCAGCCAAAGCCCCCAACCCCTGACA 765
 766 TTCTCGAGATCGAGTTCAAAAAAGGGTCCCTGTGAAGGTGACCA 810
 811 ACGTCAAGGATGGCACCACCACCAGACCTCCTTGAGCTCTTCA 855
 856 TGTACCTGAACGAAGTCGCGGGCAAGCATGGCGTGGGCCGTATTG 900
 901 ACATCGTGGAGAACCGCTTCATTGGAATGAAGTCCCAGGTATCT 945
 946 ACGAGACCCAGCAGGCACCATCCCTTTACCATGCTCATTTAGACA 990
 991 TCGAGGCCTTCACCATGGACCGGGAAGTGCGCAAAATCAAACAAG 1035
 1036 GCCTGGGCTTGAAATTTGCTGAGCTGGTGTATACCGGTTTCTGGC 1080
 1081 ACAGCCCTGAGTGTGAATTTGTCCGCCACTGCATCGCCAAGTCCC 1125
 1126 AGGAGCGAGTGAAGGGAAAGTGCAGGTGTCCGTCTCAAGGGCC 1170
 1171 AGGTGTACATCCTCGGCCGGGAGTCCCCACTGTCTCTCTACAATG 1215
 1216 AGGAGCTGGTGAGCATGAACGTGCAGGGTGATTATGAGCCAAC TG 1260
 1261 ATGCCACCGGTTTCATCAACATCAATTCCCTCAGGCTGAAGGAAT 1305

1306 ATCATCGTCTCCAGAGCAAGGTCACTGCCAAATAGACCCGTGTAC 1350
1351 AATGAGGAGCTGGGGCCTCCTCAATTTGCAGATCCCCAAGTACA 1395
1396 GGCGCTAATTGTTGTGATAATTTGTAATTGTGACTTGTTCCTCCC 1440
1441 GGCTGGCAGCGTAGTGGGGCTGCCAGGCCCCAGCTTTGTTCCCTG 1485
1486 GTCCCCCTGAAGCCTGCAAACGTTGTCATCGAAGGGAAGGGTGGG 1530
1531 GGGCAGCTGCGGTGGGGAGCTATAAAAATGACAATTTAAAAGAGAC 1575
1576 ACTAGTCTTTTATTT 1590

ANEXO II. Secuencia de cDNA predicha para chimpancé

>cDNA ASS Pan

1 CTGCTCTGCCGCTGCCACCGCTGCCCCGAGCCCGAGTGGTCCACTGCACTGTGAAGACAG 60
 61 ATTCCAGACGCCGGAACTCACGCCCTCCAATCCCAGACGCTATGTCCAGCAAAGGCTCC 120
 121 GTGGTTCTGGCCTACAGTGGCGGCCTGGACACCTCGTGCATCCTCGTGTGGCTGAAGGAA 180
 181 CAAGGCTATGACGTCATTGCCTATCTGGCCAACATTGGCCAGAAGGAAGACTTCGAGGAA 240
 241 GCCAGGAAGAAGGCACTGAAGCTTGGGGCCAAAAAGGTGTTTCATTGAGGATGTCAGCAAG 300
 301 GAGTTTGTGGAGGAGTTCATCTGGCCGGCCATCCAGTCCAGCGCACTGTATGAGGACCGC 360
 361 TACCTCTGGGCACCTCTCTTGCCAGGCCCTGCATCGCCCGCAAACAAGTGGAAATCGCC 420
 421 CAGCGGGAGGGGGCCAAGTATGTGTCCACGGCGCCACAGGAAAGGGGAACGATCAGGTC 480
 481 CGGTTTGAGCTCAGCTGCTACTCGCTGGCCCCCACATAAAGGTCATTGCTCCCTGGAGG 540
 541 ATGCCCCGAATTCTACAACCGGTTCAAGGGCCGCAATGACCTAATGGAGTACGCAAAGCAA 600
 601 CACGGGATTCCCATCCCAGTCACTCCCAAGAACCCGTGGAGCATGGATGAGAACCTCATG 660
 661 CACATCAGCTACGAGGCTGGAATCCTGGAGAACCCCAAGAACCAAGCGCCTCCAGGTCTC 720
 721 TACACGAAGACCCAGGACCCAGCCAAAGCCCCAACACCCCTGACATTCTCGAGATCGAG 780
 781 TTCAAAAAAGGGGTCCCTGTGAAGGTGACCAACGTCAAGGATGGCACCACCCACCAGACC 840
 841 TCCTTGAGGCTCTTCATGTACCTGAACGAAGTCGCGGGCAAGCATGGCGTGGGCCGTATT 900
 901 GACATCGTGGAGAACCGCTTCATGGAATNNNNNNNNNNNGTATCTACGAGACCCCAGCA 960
 961 GGCACCATCCTTTACCATGCTCATTTAGACATCGAGGCCTTACCATGGACCGGGAAGTG 1020
 1021 CGCAAAATCAAACAAGGCCTGGGCTTGAAATTTGCTGAGCTGGTGTATAACCGTTTCTGG 1080
 1081 CACAGCCCTGAGTGTGAATTTGTCCGCCACTGCATCGCCAAGTCCCAGGAGCGAGTGGAA 1140
 1141 GGGAAAAGTGCAGGTGTCCGTCCCTCAAGGGCCAGGTGTACATCCTCGGCCGGGAGTCCCCA 1200
 1200 CTGTCTCTTACAATGAGGAGCTGGTGAGCATGAACGTGCAGGGTGATTATGAGCCAACT 1260
 1261 GATGCCACCGGGTTCATCAACATCAATTCCTCAGGCTGAAGGAATATCATCGTCTCCAG 1320
 1321 AGCAAGGTCCTGCCAAATAGACCCCTGTACAATGAGGAGCTGGGGCCTCCTCAATTTGC 1380
 1381 AGATCCCCCAAGTACAGGCGCTAATTGTTGTGATAATTTGTAATTGTGACTTGTCTCCC 1440
 1441 CGGCTGGCAGTGTAGTGGGGCTGCCAGGCCCCAGCTTTGTTCCCTGGTCCCCCTGAAGCC 1500
 1501 TGCAAACGTTGTCATCGAAGGGAAGGGTGGGGGGGGGGCAGCTGCGGCGGGGAGCTATA 1560
 1561 AAAATGACAATTAAGAGACACTAGTCTTTTGTCTTCT 1598

ANEXO III. Polipéptido predicho para chimpancé

>ASS Prot Chimpancé

```
1  MSSKGSVVLAYSGLDTSCILVWLKEQGYDVIAYLANIGQKEDFEEARKKALKLGAKKVF  60
61  IEDVSKEFVEEFIWPAIQSSALYEDRYLLGTSLARPCIARKQVEIAQREGAKYVSHGATG 120
121 KGNDQVRFELSCYSLAPHIKVIAPWRMPEFYNRFKGRNDLMEYAKQHGIPIPVTPKNPWS 180
181 MDENLMHISYEAGILENPKNQAPPGLYTKTQDPAKAPNTPDILEIEFKKGVKVTNVKD  240
241 GTTHQTSLELEFMYLNEVAGKHGVGRIDIVENRFMEXXXIYETPAGTILYHAHLDEAF  300
301 TMDREVRKIKQGLGLKFAELVYTGFWHSPECFVRHCIAKSQERVEGKVQVSVLKGQVYI  360
361 LGRESPLSLYNEELVSMNVQGDIYPTDATGFININSLRLKEYHRLQSKVTAK  412
```

ANEXO IV. Polipéptido predicho para la retrosecuencia de ratón

>ASS Prot Ratón

```
1  MSSKGSVVLAYSGLDTSILVWLKEQGYDVIAYLANIGQKEDFEEARKKALKLGAKKVF  60
61  IEDVSKEFVEEFIWPAVQSSALYEDRYLLGTSLARPCIARRQVEIAQREGAKYVSHGATG 120
121 KGNDQVRFEELTCYSLAPQIKVIAPWRMPEFYNRFKGRNDLMEYAKQHGIPIPVSPKSPWS 180
181 MDENLMHISYEAGILENPKNQAPPGLYTKTQDPAKAPNTPDVLEIEFKKGVKVTNIKD 240
241 GTTRTTSLELEFMYLNEVAGKHGVGRIDIVENRFIGMKSRIYETPAGTILYHAHLDEAF 300
301 TMDREVRKIKQGLGLKFAELVYTGFWHSPECEFVRHCIQKSQERVEGKVQVSVFKGQVYI 360
361 LGRESPLSLYNEELVSMNVQGDYEPIDATGFININSLRLKEYHRLQSKVTAK 412
```

Bibliografía

Alcaraz-Peraza, L. (2005) Retropseudogenes de las chaperoninas mitocondriales en el genoma de *Mus musculus* como modelo para estudiar la expresión genética en el pasado. Tesis de licenciatura en Biología. Facultad de Ciencias, UNAM. México.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Barberán-Soler, S. (2002) Retropseudogenes humanos un análisis genómico y evolutivo. Tesis de licenciatura en Biología. Facultad de Ciencias, UNAM. México

Bredt, D.S., P.M. Hwang and S.H. Snyder (1990) Localization of nitric oxide synthase indicating a neutral role for nitric oxide. *Nature* 347: 768-770.

Bromham, L., M.J. Phillips and D. Penny (1999) Growing up with dinosaurs: molecular dates and the mammalian radiation. *Trends Ecol. Evol.* 14: 113-118.

Burnett, A., D.D. Ricker, S.L. Chamness, M.P. Maguire, J.K. Crone, D.S. Bredt, S.H. Snyder and T.S.K. Chang (1995) Localization of nitric oxide synthase in the reproductive organs of the male rat. *Biol. Reprod.* 52: 1-7.

D'Errico, I., G. Gadaleta and C. Saccone (2004) Pseudogenes in metazoa: Origin and features. *Briefings in functional genomics and proteomics.* 3 (2): 157-167.

Deininger, P.L and M.A. Batzer (2002) Mammalian retroelements. *Genom. Research* 12:1455-1465.

Devor, E.J. and K.A. Moffat-Wilson (2003) Molecular and temporal characteristics of human retropseudogenes. *Human Biology.* 75(5): 661-672.

Ebersberger, I., D. Metzler, C. Schwarz and S. Pääbo (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am.J.Hum.Genet.*70: 1490-1497.

Esnault, C., J. Maestre and T. Heidmann (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics* 24: 363-367.

Feuk, L., J.R. MacDonald, T. Tang, A.R. Carson, M. Li, G. Rao, R. Khaja and S.W. Scherer (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics* 1(4): 0489-0498.

Fletcher, B.H, A. Cassady, K.M. Summers, A. Cavanagh (2001) The murine chaperonin 10 gene family contains an intronless, putative gene for early pregnancy factor, Cpn10-rs1. *Mammalian Genome* 12: 133-140.

Freytag, S.O., H.G.O. Bock, A.L. Beaudet and W.E. O'Brien (1984) Molecular structure of the human Arginosuccinate synthetase pseudogenes. *J. Bio. Chem.* 259(5): 3160-3166.

Freytag, S.O., A.L. Beaudet, H.G.O. Bock and W.E. O'Brien (1984) Molecular structure of the human Arginosuccinate synthetase gene: Occurrence of alternative mRNA splicing. *Mol. Cel. Biol.* 4: 1978-1984.

Ghose, C. and F.M. Raushel (1985) Determination of the mechanism of the argininosuccinate synthetase reaction by static and dynamic quench experiments. *Biochemistry* 24(21): 5894-5898.

Gonçalves, I., L. Duret and D. Mouchiroud (2000) Nature and structure of human genes than generate retropseudogenes. *Genome* 10: 672-678.

Graur, D., Y. Shuali and W.H. Li (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J. Mol. Evol.* 28: 279-285.

Han, K., S.K. Sen, J. Wang, P.A. Callinan, J. Lee, R. Cordaux, P. Liang and M. Batzer (2005) Genomic rearrangements by LINE-1 insertion-mediated deletion in human and chimpanzee lineages. *Nucleic Acid Research* 33(13): 4040-4052.

Hardison, R. (1998) Hemoglobins from bacteria to man: Evolution of different patterns of gene expression. *J. Exp. Biol.* 201: 1099-1117.

Higgins D., J. Thompson, T. Gibson, J.D. Thompson, D.G. Higgins and T.J. Gibson (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.

Hurteau, G.J. and S.D. Spivack (2002) mRNA-specific reverse transcription-polymerase chain reaction from human tissue extracts. *Anal. Bioch.* 307: 304-315.

Husson, A., C. Basse-Lagnel, A. Fairand, S. Renouf and A. Lavoigne (2003) Argininosuccinate synthetase from the urea cycle to citrulline-NO cycle [review]. *Eur. J. Biochem* 270, 1887-1899.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.

Kazazian, H.H. (2004) Mobile elements: Drivers of genome evolution. *Science* 303: 1626-1632.

Kim, H., C. Moon, M. Ahn, Y. Lee, H. Kim, S. Kim, T. Ha, Y. Jee and T. Shin (2005) Expression of nitric oxide synthase isoforms in the porcine ovary during follicular development. *J.V.Sci.* 6(2): 97-101.

Kimura, M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111-120.

Knowles, R.G. and S. Moncada (1994) Nitric oxide synthases in mammals. *Biochem. J* 298: 249-258.

Kobayashi, K., M.J. Jackson, D.B. Tick, W.E. O'Brien and A.L. Beaudet (1990) Heterogeneity of mutation in argininosuccinate synthetase causing human citrullinemia. *J. Bio. Chem.* 265(19): 11361-11367.

Korenberg, J.R. and M.C. Rykowski (1988) Human genome organization: Alu, LINE and the molecular structure of metaphase chromosome bands. *Cell* 53: 391-400.

Kumar S., K. Tamura and M. Nei (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics* 5: 150-163.

Lemke, C.T. and P.L. Howell (2001) The 1.6 Å crystal structure of *E. coli* argininosuccinate synthetase suggests a conformational change during catalysis. *Structure.* 9: 1153-1164.

Lemke, C.T. and P.L. Howell (2002) Substrate induced conformation changes in Argininosuccinate synthetase. *J. Biol. Chem.* 277 (15): 13074-13081.

Li, W.H. and D. Graur (2000) *Fundamental of Molecular Evolution.* Sinauer Associates, Inc. Sunderland Massachusetts.

Lissbrant, E., U. Löfman, O. Collin and A. Bergh (1997) Is nitric oxide involved in the regulation of the rat testicular vasculature? *Biolo. Reprod.* 56: 1221-1227.

Long, M., M. Deutsch, W. Wang, E. Betrán, F.G. Brunet and J. Zhang (2003) Origin of new genes: evidence from experimental and computational analyses. *Genetica* 118: 171-182.

Martín-Hernández, I. (2005) Una aproximación a los desórdenes hereditarios del ciclo de la urea en el hombre [Revisión]. *Rev. Biomed.* 16: 193-206.

Mighell, A.J., N.R. Smith, P.A. Robinson, and A.F. Markham (2000) Vertebrate pseudogenes [Minireview]. *FEBS Letters* 468: 109-114.

Mori, M., T. Gotoh, A. Nagasaki, M. Takiguchi and T. Sonoki (1998) Regulation of the urea cycle enzyme genes in nitric oxide synthesis. *J. Inher. Metab. Dis.* 21 (Suppl. 1): 59-71.

Morris, S.M. (2002) Regulation of enzymes of the Urea Cycle and Arginine metabolism. *Annu. Rev. Nutr.* 22: 87-105.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.

Nei, M. and S. Kumar (2000) *Molecular evolution and phylogenetics*. Oxford University Press, New York.

Niimura, Y. and M. Nei (2005) Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene* 346: 23-28.

Olson M.V. and A. Varki (2003) Sequencing the chimpanzee genome: Insights into human evolution and disease [review]. *Nature Genetics* 4: 20-28.

Park, C.S., H.S. Lee, H.Y. Lee and G. Krishna (1997) An unprocessed pseudogene of Nitric Oxide Synthase gene in human. *Nitric Oxide: Biology and Chemistry* 1(4): 294-300.

Pesce, A., M. Bolognesi, A. Bocedi, P. Ascenzi, S. Dewilde, L. Monees, T. Hankeln and T. Burmester (2002) Neuroglobin and cytoglobin: Fresh blood for the vertebrate globin family. *EMBO reports* 3(12): 1146-1151.

Pendleton, L.C., B.L. Goodwin, B.R. Flam, L.P. Solomonson and D.C. Eichler (2002) Endothelial Arginosuccinate synthetase mRNA 5'-Untranslated region diversity. *J. Bio. Chem.* 277 (28): 25363-25369.

Ross, M.T., D.V. Grafham, A.J. Coffey *et al* (2005) The DNA sequence of the human X chromosome. *Nature* 434: 325-337.

Roy-Engel, A., M.L. Carroll, M. El-Sawy, A.H. Salem, R.K. Garber, S.V. Nguyen, P.L. Deininger and M.A. Batzer (2002) Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* 316: 1033-1040.

Rutherford K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.A. Rajandream and B. Barrell (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16 (10): 944-945.

Saitou, N. and M. Nei (1987) The Neighbor-Joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

Samonte, R.V. and E.E. Eichler (2002) Segmental duplication and the evolution of the primate genome. *Nature Reviews: Genetics* 3: 65-72.

Soares, M. B., E. Schon, A. Henderson, S.K. Karathabasis *et al* (1985) RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. *Mol. Cell. Biol.* 5: 2090–2103.

Steinemann, S.S. and M. Steinemann (2005) Y chromosomes: born to be destroyed. *BioEssays* 27: 1076-1083.

Stryer, L. (1995) *Bioquímica*. Reverte. España.

Teixeira, D., L. Santaolara y E. Alonso (2003) La arginina en su contexto metabólico y fisiológico. *Acta Bioquím. Clín. Latinoamericana* 37 (2): 165-179.

Telen M. Eritrocitos maduros. En: Lee GR, TC. Bithell, J. Foerster, J. Athens y J. Lukens (eds.) (1994) *Wintrobe: Hematología clínica*. Intermédica. Buenos Aires.

Varki, A. (2000) A chimpanzee genome project is a biomedical imperative. *Genome Res.* 10: 1065–1070.

Valdés-López, V., A. Vilchis-Peluyera, L. Alba-Lois, B. Rodarte-Mugia, C. Segal-Kischinevzky, B. Rodríguez-Ponce y L.D. Alcaraz-Peraza (2004) Evolución genómica: aparición y divergencia de retropseudogenes. *Sci. Amer. Lat.* 22: 32-33.

Vilchis-Peluyera, A.J. (2006) Análisis evolutivo de retropseudogenes de las familias de proteínas de choque térmico HSP10 y HSP60 en el genoma humano. Tesis de Maestría en Ciencias Biológicas. Facultad de Ciencias, UNAM. México.

Wang, W., J. Zhang, C. Alvarez, A. Llopart and M. Long (2000) The origin of the *Jingwei* gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. *Mol. Biol. Evol.* 17: 1294–1301.

Wine, R.N., W.W. Ku, L.H. Li, and R.E. Chapin (1997) Cyclophilin A is present in rat germ cells and is associated with spermatocyte apoptosis. *Reproductive Toxicology Group. Biol. Reprod.* 56: 439–446.

Woischnik, M. and C. Morales (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome* 12: 885-893.

Wu, G. and S.M. Morris (1998) Arginine metabolism: nitric oxide and beyond [Review Article] *Biochem. J.* 336: 1-17.

Zhang, Z. and M. Gerstein (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferres from pseudogenes. *Nucleic Acids Research.* 31 (18): 5338-5348.

Zhang, Z., P.M. Harrison, Y. Liu and M. Gerstein (2003) Millions of years of evolution preserved: a comprehensive catalog of processed pseudogenes in the human genome. *Genome* 13: 2541-2558.