UNIVERSIDAD DE CHILE FACULTAD DE CIENCIAS ESCUELA DE PREGRADO

"Búsqueda de SNPs provenientes de regiones duplicadas en el genoma de un único individuo de Salmón del Atlántico (*Salmo salar*)"

Seminario de Título entregado a la Universidad de Chile en cumplimiento parcial de los requisitos para optar al Titulo de Ingeniero en Biotecnología Molecular.

LUIS ROBERTO ZAPATA ORTIZ

Dra. Patricia Iturra
Directora de Seminario de Título:

Dr. Alejandro Maass
Co- Director de Seminario de Título

Junio de 2010
Santiago - Chile

ESCUELA DE PREGRADO – FACULTAD DE CIENCIAS – UNIVERSIDAD DE CHILE

INFORME DE APROBACIÓN SEMINARIO DE TITULO

Se informa a la Escuela de Pregrado de la Facultad de Ciencias, de la Universidad de Chile que el Seminario de Título, presentado por el Sr.

"LUIS ROBERTO ZAPATA ORTIZ"

Búsqueda de SNPs provenientes de regiones duplicadas en el genoma de un único individuo de Salmón del Atlántico (*Salmo salar*)

Ha sido aprobado por la Comisión de Evaluación, en cumplimiento parcial de los requisitos para optar al Título de .....

Dra. Patricia Iturra
Directora Seminario de Título

Dr. Alejandro Maass
Co-Director

Comisión de Evaluación

Dr. Ricardo Cabrera
Presidente (a) Comisión

Dr. Mauricio González
Evaluador

Santiago de Chile, Junio 2010

A mi familia por su amor incondicional, en especial a mi madre que, con su confianza plena y esfuerzo incansable, me entregó las herramientas más valiosas para ser feliz día a día.

Quiero expresar mi profundo agradecimiento:

A mi directora de Tesis, Dra. Patricia Iturra, por su confianza y estímulo para crecer intelectualmente y llevar a cabo este proyecto.

A mi co-director de Tesis, Dr. Alejandro Mass, por entregarme la confianza y el espacio necesarios para desarrollar esta memoria.

Al Dr. Juan Medrano por darme la oportunidad de trabajar en su laboratorio en Davis durante mi periodo de intercambio y luego para hacer los experimentos de mi tesis. Por guiarme y ayudarme a que mis ideas se hicieran realidad.

Al Dr. Willie Davidson por facilitarme las secuencias de DNA que hicieron posible este trabajo.

A mis compañeros de laboratorio de Bioinformática, Alex, Cristian, Ma. Paz, Marko, Felipe, Raúl y Andrés, que siempre estuvieron dispuestos a ayudarme a resolver mis dudas informáticas.

Al proyecto CORFO-INNOVA 07CN13PBT-41, FONDAP y BASAL-CMM que alimentaron mis neuronas durante la realización de esta tesis.

A mis padres por su ejemplo de esfuerzo y su valioso apoyo durante mi época de estudiante.

A mis amigos por estar ahí siempre.

A mi amada por ser el pilar de mi felicidad y mi compañera de vida.

Gracias.

Abstract

In recent years Single Nucleotide polymorphisms (SNPs) have become the most popular molecular marker tool used in genetics. *In silico* SNP discovery strategies have been focused on the alignment of input fragments covering the same region of the genome whose sequence differ due to polymorphic variation in the population. Genomic duplications represent a biased source of SNPs. The nucleotides that are present in equivalent locations of the duplicated regions, but vary between the two loci, would appear to have a heterozygous genotype by coamplification of these duplicates. The SNPs found in duplicons are known as paralogous squence variants (PSVs). If one of these SNPs is also polymorphic in one of the duplicated copies, it is known as multi site variation (MSV). These markers are not appropiate to perform mapping or asociation studies.

The common ancestor of salmonids experienced a whole genome duplication event recently (100 MYA). In the present study, a pipeline was developed in order to predict duplicated regions within an individual *Salmo salar* genome using SNP data and repetitive element masking. 217000 BAC end sequences coming from a single individual genome were assembled. There were 4991 SNPs in 1352 contigs. At the final stage of the pipeline, the final set of candidate duplicated contigs was reduced to 956. 10 regions with a polymorphism closely linked to a gene were selected to genotype three different individuals. Each SNP was validated in 6 out of 9 succesfully sequenced regions. One contig corresponding to 227 bp of *Hsp70* gene CDS revealed 9 SNPs. This locus is duplicated and the predicted SNPs correspond to PSVs instead of true SNPs. PSVs and MSVs have to be considered behind a new concept in genetics, defined as intragenomic variation.

The use of this pipeline combined with SNP discovery in multiple individuals will allow researchers to asess the true nature of the SNPs which they are working with.

Resumen

El genoma del salmón acumula cuatro rondas de duplicación genómica en su historia evolutiva: un evento específico de salmones, otro específico del linaje de los teleósteos y dos rondas ocurridas en vertebrados hace 600 MYA (Hipótesis 2R). La búsqueda de marcadores en esta especie es necesaria para realizar estudios evolutivos, de mapeo o de asociación a genes de interés, por lo que la presencia de regiones duplicadas impone un problema al momento de establecer el origen de estos marcadores.

En el presente estudio se implementó una estrategia bioinformática donde 217 000 secuencias génomicas provenientes de un único individuo fueron ensambladas y analizadas en busca de SNPs. Se encontraron 4 900 SNPs distribuidos en 1350 contigs de un total de 25 000. Luego, se filtró un 32% de secuencia repetitiva para descartar los polimorfismos presentes en estos elementos llegando a un set final aproximado de 2800 SNPs en 950 contigs. Todas estas secuencias fueron anotadas con la información disponible en las bases de datos públicas.

Se estableció la presencia de loci duplicados básandose en la hipótesis que los SNPs encontrados en un único individuo corresponden a diferencias nucleotídicas entre regiones parálogas. La preservación de estas regiones está sujeta a la adquisición de mutaciones que le confieran una ventaja selectiva, ya sea en una región regulatoria o codificante, que les permita escapar a la selección purificadora de la evolución.

Se seleccionaron 10 contigs para ser secuenciados y genotipar los polimorfismos putativos en tres individuos. Se encontró evidencia que seis de estos SNPs corresponden a variantes dadas por la duplicación. Las cuatro regiones restantes presentaron alguna dificultad en el genotipado que puede ser corregida y se discute en el texto. Una de las regiones, correspondiente al gen *Hsp70*, presentó un patrón de alta densidad de SNPs relacionado con la acumulación de mutaciones neutrales en la secuencia de dos copias

parálogas divergentes de un mismo gen.

La estrategia propuesta permite la búsqueda de SNPs en secuencias provenientes de un sólo individuo, la identificación de regiones parálogas y la generación de una base de datos genómica. Esta estrategia puede ser combinada con la búsqueda de SNPs en múlitples individuos, para así diferenciar si estos marcadores provienen de regiones duplicadas o no. Además, puede ser aplicada a cualquier especie siendo probable que su efectividad aumente en organismos que hayan sufrido una duplicación reciente, como el salmón, ya que se encuentra una mayor cantidad de loci funcionales que han adquirido mutaciones en sus copias y se han preservado como duplicados.

Índice general

Índice de figuras

Índice de cuadros

1. Introducción

1.1. Preámbulo

El salmón del Atlántico (*Salmo salar*) es una de las especies acuícolas más importantes para nuestro país, siendo Chile uno de los mayores productores mundiales de esta especie en cautiverio junto con Canadá y Noruega. El salmón es una especie de gran impacto en áreas como la pesca deportiva y la alimentación. Numerosos proyectos han aumentado la información biológica de las especies salmonídeas, poniendo a disposición pública datos relativos a su fisiología, genética, ecología y desarrollo, entre otros (Thorgaard y col., 2002).

Gracias a los recientes proyectos de secuenciación a gran escala se dispone actualmente de múltiples secuencias de **ESTs** (Thorsen y col., 2005; Adzhubei y col., 2007), un mapa físico (Ng y col., 2005), varios mapas genéticos (Hayes y col., 2007b; Moen y col., 2004; Gilbey y col., 2004) y una biblioteca BAC con las secuencias de los extremos de los clones (BAC End sequences o **BES**) (Thorsen y col., 2005). Sin embargo, la cantidad de elementos repetitivos y el estado duplicado de su genoma impone una dificultad importante al momento de estudiar su genoma y buscar marcadores moleculares, tales como SNPs o microsatélites, en genes de interés.

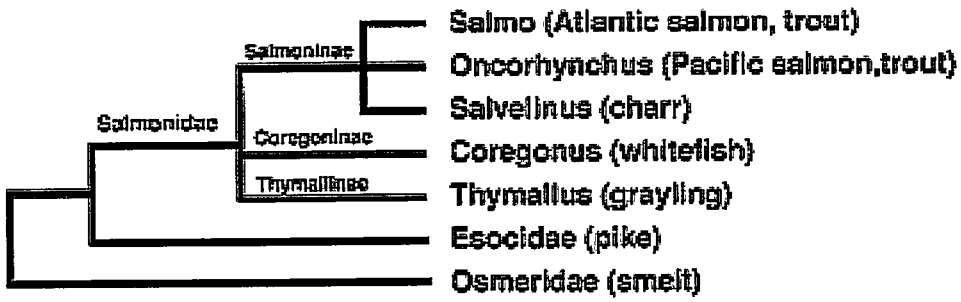

Figura 1.1: Relaciones filogenéticas entre las principales clases de salmónidos y sus géneros cercanos. ((Nelson, 2006)).

1.2. Salmónidos

1.2.1. Introducción

La familia Salmonidae está compuesta por tres subfamilias, Coregoninae, Thymallinae y Salmoninae. Esta última corresponde a la subfamilia que alberga a los géneros, *Brachymystax, Hucho, Salmothymus, Salvelinus, Salmo* y *Oncorhynchus* (revisado en Phillips & Rab (2001)). Dentro de ésta familia existe acuerdo que *Brachymystax* y *Hucho* están más relacionados a *Salvelinus* y que *Onchorhynchus y Salmo* son más cercanos entre sí (Fig. 1.1).

Estas especies tienen una distribución holártica a través de gran parte de Eurasia y Norteamérica. Muchas de ellas se pueden encontrar a lo largo de estas zonas presentando una gran variedad genotípica y fenotípica tanto entre como dentro de las distintas poblaciones. Esta variación es relativamente reciente y se observa en rasgos tales como la edad de maduración, la frecuencia de reproducción, el tamaño del cuerpo, el dimorfismo sexual, la temporada de cruza y su morfología, entre otros (Willson, 1997).

Todas las especies de la familia Salmonidae descienden de un ancestro común que se cree sufrió un evento de tetraploidización hace 25-100 millones de años atrás luego de la

expansión colonizadora de los teleósteos (Allendorf & Thorgaard, 1984). Los salmónidos son especies que están reestableciendo la diploidía y en el cual aproximadamente el 50 % de sus loci se mantiene como duplicados funcionales (Bailey y col., 1978). La rediploidización es el proceso que da cuenta de los polimorfismos cromosómicos observados en los salmones. Su número diploide (2n) varía de 52 a 102 cromosomas y el número de brazos cromosómicos (**NF**) varía de 72 a 170. Sin embargo, la mayoría de las especies en cada género presenta un NF cercano a 100, lo que sugiere que el ancestro común de los salmones poseía un cariotipo compuesto por 48-50 cromosomas telocéntricos que durante un evento de autopoliploidización duplicó su número para luego sufrir una serie de rearreglos cromosómicos que redujeron o aumentaron su 2n y generaron la variación observada actualmente. Esto corresponde a una de las evidencias más sólidas que respalda el evento de WGD ocurrido en los salmónidos (Phillips & Rab, 2001).

Los peces Teleósteos, que representan a la mitad de las especies de vertebrados en el mundo y en el cual se encuentran los salmónidos, sufrieron un evento de duplicación de su genoma hace aproximadamente 350 millones de años (Hoegg y col., 2004; Volff, 2005). En conjunto con dos rondas de duplicaciones sufridas por los vertebrados (hipótesis 2R) (Taylor & Raes, 2004), el genoma de los salmónidos acumularía 4 duplicaciones genómicas (cuatro WGD o 4R) (Davidson & Koop, 2008). Esto convierte al salmón del Atlántico en una especie modelo para evaluar teorías respecto a duplicación genómica, divergencia, especiación y evolución molecular.

1.2.2. Salmón del Atlántico

Historia

El salmón del Atlántico es una de las especies mejor estudiadas de los salmónidos. Numerosos estudios han sido publicados relativos a su ecología, fisiología, genética y distribución (Thorgaard y col., 2002). Su rango de distribución nativa se encuentra en el sector atlántico de Europa, principalmente en forma no anádroma en Noruega y Suecia, y a lo largo de la región de Karelia en Rusia y el mar Báltico. En Norteamérica estas formas se encuentran en el rio Hudson y en el lago Ontario hasta el norte de Quebec. También están presentes en Groenlandia y algunas regiones de Europa como Portugal y España pero en menor cantidad. La mayor cantidad de las formas anádromas del salmón se encuentran extintas en muchos ríos (MacCrimmon & Gots, 1979).

Los primeros intentos de reproducir artificialmente las diversas especies de Salmón se remontan al siglo XIV en Francia. Luego, la fertilización artificial de huevos de trucha y salmón en el siglo XVIII permitió el establecimiento de las primeras pisciculturas de salmón en Francia y Estados Unidos. Luego, otros países como Inglaterra, Japón y Alemania, lideraron los esfuerzos por realizar reproducción artificial para aumentar la producción de estas especies a lo largo de sus colonias. De esta forma, enviaron huevos fertilizados a distintas partes del mundo incluyendo Australia y Nueva Zelanda, para despúes introducirlos en Sudamérica. En Chile la especie fue introducida desde la piscicultura del río McCloud en Estados Unidos.

Con las mejoras técnicas en el cultivo a mitad del siglo XX, la industria creció enormemente relegando la captura de salmón nativo a niveles bajos (actualmente es entre 300-600 veces menor que las prácticas de acuicultura de salmón cultivado). A partir de la década del 70 en adelante la producción mundial ha crecido a un nivel de más de 1 millón de toneladas anuales, siendo Noruega y Chile los que producen casi un 75% del

total, seguidos por Canadá e Inglaterra. El salmón del Atlántico es la especie salmonídea de mayor producción seguido por Trucha arcoiris y salmón Coho (Araneda y col., 2008).

Biología

El salmón del Atlántico es una especie salmónida de regiones temperadas y subárticas del norte del oceáno Atlántico que pueden encontrarse en ríos o mares de agua salada. Sin embargo, las intensas prácticas de cultivo alrededor del mundo han provocado que la mayoría sólo pueda encontrarse en su forma no anádroma. Generalmente, la hembra desova alrededor de 2000 - 15000 huevos y no muere como otras especies de salmones, por lo que es común que vuelvan al mismo río a desovar. Presentan fertilización externa con la hembra protegiendo los huevos sobre pequeños nichos de roca cercanos a la desembocadura de los ríos. En esta fase existe una mortalidad natural del 70 %. Luego de pasar uno o dos años en el río se presentan los cambios morfológicos de la smoltificación, donde las formas anádromas migran al mar para pasar entre 1 y 4 años y luego regresar para el período de desove. Las formas residentes de los ríos maduran entre los 2 y 5 años, viven alrededor de 10 o más y desovan consecutivamente año a año. En la fase adulta se alimentan de peces y crustáceos, llegan a medir entre 50-120 cm y pesar entre 2-30 kg (Webb y col., 2006).

1.2.3. Genoma del salmón

La evidencia respecto al origen autopoliploide del genoma del salmón se encuentra bien documentada: 1) Al comparar secuencias de DNA idénticas provenientes de brazos cromosómicos distintos se observa que alrededor de un 50 % de los genes presentes en éstas se expresa activamente (Nichols y col., 2003). 2) Durante la meiosis en machos se observan configuraciones cuadrivalentes que involucran cromosomas metacéntricos en

conjunto con algún par acrocéntrico (Allendorf & Thorgaard, 1984). 3) Presentan tasas de segregación que indican herencia tetrasómica parcial o completa luego de la meiosis 4) Un número de brazos cromosómicos NF=100 cuando la mayoría de los peces tele ósteos presenta la mitad de este valor (Phillips & Rab, 2001). 5) Para el salmón del Atlántico (y la trucha arcoiris) se ha descrito la presencia de 14 cluster Hox en distintas localizaciones cromosómicas, lo que se relaciona con la teoría 4R que aceptaría hasta 16 copias (1 - 2 - 4 - 8 - 16).

Danzmann y col. (2008) concluyeron que dos brazos cromosómicos en *D. rerio* y *O. latipes* pueden, en la mayoría de las instancias, relacionarse con al menos 4 brazos cromosómicos de las especies salmónidas. Basándose en que la mayoría de los teleósteos actuales posee un genoma haploide de 24-25 cromosomas se ha propuesto que el ancestro diploide de los salmones poseía 48 cromosomas telocéntricos. Luego de la duplicación, el primer salmón tetraploide aumentó al doble este número quedando con un NF cercano a 100. Los peces más cercanos a los salmones son los osmeriformes que poseen un número diploide entre 50-56 lo que apoya que el tiempo de duplicación ocurrió después de la separación de esta especie con los salmónidos (ver fig. 1.1).

El salmón del Atlántico presenta un genoma que posee entre 54-58 cromosomas y un NF entre 72-76 bastante lejano a 100. La presencia de bandas C intersticiales en los cromosomas telocéntricos más largos indica que su cariotipo actual se formó principalmente a partir de múltiples fusiones de tipo tandem seguido de inversiones cromosómicas. Sin embargo, al contar dos brazos por cada cromosoma su NF resulta ser 100 (Phillips & Rab, 2001), al igual que el resto de los salmónidos. La compleja arquitectura genómica de los peces, y en especial los salmónidos, les ha permitido una adaptación y especiación rápida en respuesta a diversos medio ambientes (Amores y col., 1998). Esta capacidad adquirida es consecuencia de la evolución basada en un

modelo de mutación-duplicación-selección.

La información genómica disponible de esta especie ha ido en un aumento considerable en los últimos años. Entre ésta, se encuentran las secuencias génomicas de los extremos de los clones de una librería BAC (BAC End sequences o BES) realizada a partir de la construcción de un mapa físico proveniente de la digestión parcial del genoma de un único individuo de *Salmo salar* con la enzima de restricción HindIII. Estos extremos flanquean alrededor de 110 000 clones, tienen un largo promedio de 900 pares de bases y la distancia media entre ambos extremos de cada clon es de aproximadamente 180 Kb (Thorsen y col., 2005). La disponibilidad de estas secuencias provenientes de un único individuo permite la búsqueda de polimorfismos y marcadores a lo largo de múltiples regiones genómicas.

La presencia de regiones duplicadas no identificadas a lo largo del genoma del salmón ha generado una dificultad constante en el descubrimiento de marcadores moleculares, en particular SNPs. En estos loci existe una presión selectiva menor debido a la presencia previa de un locus funcional (Gut & Lathrop, 2004), por lo que cualquier cambio de base en su secuencia puede permanecer en esa región generando una mayor densidad de mutaciones (Hayes y col., 2007a; Rise y col., 2004).

Elementos transponibles

Los elementos transponibles (TE) son secuencias de DNA que contienen en ella todo lo necesario para cortarse a sí mismos y moverse a distintos lugares del genoma, lo que puede traer severas consecuencias ya que se pueden insertar en secuencias codificantes, interrumpiendo la secuencia aminoacídica correcta, en regiones regulatorias modificando la expresión genica o promoviendo rearreglos cromosómicos debido a la naturaleza repetitiva de su secuencia (Feschotte, 2008). Sin embargo, estos elementos

no siempre generan efectos adversos, ya que su actividad mutacional puede contribuir a la diversidad genética del organismo constituyendo también una fuente de innovación genómica para las especies (Kazazian, 2004).

Los TEs presentes en peces tienen una mayor diversidad y presencia en comparación con el humano y se localizan especialmente en áreas heterocromáticas. El genoma del salmón presenta aproximadamente un 30 % de elementos repetitivos que incluyen a todo tipo de transposones. Se ha observado un incremento en la actividad de estos elementos luego de eventos de alopoliploidización o hibridación interespecies. Esto ha llevado a Davidson & Koop (2008) a formular la hipótesis que durante el proceso de rediploidización en salmones, la actividad transposónica aumentó para facilitar la reestabilización de su genoma. Los autores concluyeron que el aumento en la actividad replicativa de los TEs coincidió con el tiempo de radiación del género *Salmoninae* en *Salmo, Oncorhynchus y Salvelinus* y que aún existe un intenso período de actividad que no es claro si está relacionado con el proceso de especiación o con el proceso de reestabilizar el genoma para evitar la formación de estructuras tetravalentes (De Boer y col., 2007).

1.3. Mutaciones y duplicaciones

En general existen dos grandes categorías de mutaciones descritas en la literatura, mutaciones a nivel de secuencia del DNA, donde existen pequeños cambios de bases, y mutaciones a nivel cromosómico, incluyendo las duplicaciones y deleciones. Las mutaciones pueden tener diferentes efectos y dependen del contexto de la secuencia de DNA donde se producen. Además, se ha demostrado que los mecanismos por los que ocurren estos cambios están evolutivamente bien conservados entre la mayoría de los organismos (Maki, 2002), incluyendo los peces.

Los términos mutación puntual o SNP están estrechamente relacionados ya que

el último es consecuencia del primero. Estrictamente SNP significa "Single Nucleotide Polymorphism", ó "Polimorfismo en un solo nucleótido", implica la sustitución de una única base nucleotídica (mutación) a lo largo de una secuencia de DNA por otra. Una definición más amplia involucra también deleciones o inserciones de secuencias cortas de DNA (**indels**). El término SNP se ha puesto en uso para describir una mutación puntual que está presente en al menos un 1% de una población determinada en forma de alelo. Es decir, es una mutación que se ha preservado en el genoma de alguna especie y da cuenta de la variabilidad entre individuos de una o múltiples especies. Actualmente, los SNPs se utilizan como marcadores moleculares que se heredan mendelianamente, son codominantes, preferentemente bialélicos y presentan una amplia distribución a lo largo de los genomas (Vignal y col., 2002).

La duplicación es un tipo de mutación cromosómica donde se produce una copia extra de alguna región específica, de algún cromosoma o incluso de un genoma completo. Estos duplicones o copias parálogas pueden estar adyacentes o ubicados en lugares distintos del genoma, es decir, una copia se mantiene en su región original mientras la otra puede estar en el mismo cromosoma, o en otro distinto, dependiendo del mecanismo por el que ocurrió la duplicación. Las duplicaciones se pueden clasificar como (1) duplicaciones parciales, (2) duplicaciones génicas, (3) duplicaciones segmentales de cromosomas y (4) duplicaciones del genoma completo (**WGD**, Whole Genome Duplication)(Li & Graur, 2000).

Hace casi 40 años Susumu Ohno propuso que el factor más importante en la evolución de las especies es la duplicación génica (Ohno, 1970). Estableció que la base de la evolución molecular es producto de la posibilidad de innovar manteniendo la funcionalidad original intacta. Esta posibilidad de innovación es producto del azar y sólo es ventajosa en los términos que obliga la **selección natural**. Concluyó además, que

durante un primer período evolutivo los vertebrados sufrieron dos rondas de duplicación genómica WGD (hipótesis 2R) (Ohno, 1970; Allendorf & Thorgaard, 1984).

Una de las evidencias más aceptadas de la hipótesis 2R proviene del análisis del cluster de genes Hox (Kasahara, 2007) donde el cefalocordado *Anfioxo*, posee sólo un cluster Hox y los Sarcopterigios (clase que incluye anfibios, reptiles, aves y mamíferos) presentan 4. El anfioxo es considerado un fósil viviente que representa el estado previo a la duplicación ocurrida en vertebrados. Esta razón 4:1 se ha demostrado para múltiples genes distintos (Novak y col., 2006). Especies incluidas en la clase Actinopterigios (que surgen de los sarcopterigios) como Zebrafish (*D. rerio*), Medaka (*O. latipes*) y Fugu (*T. rubripes*) presentan alrededor del doble de clusters que el humano (Amores y col., 1998) evidenciando la presencia de una tercera ronda de duplicaciones (hipótesis 3R) (Meyer & Van de Peer, 2005; Hufton y col., 2008) ocurrida antes de la radiación de los teleósteos. A su vez, el genoma de los salmónidos presenta un cuarto evento de duplicación ocurrido luego de la separación de los osmeriformes.

1.3.1. Tipos y mecanismos

Mutaciones puntuales

Durante el proceso de replicación se ha estimado que la DNA polimerasa humana sólo produce un error cada 10^8-10^{10} pares de bases copiadas, lo que implica una alta fidelidad y poca probabilidad de error (Baer y col., 2007). Además, existen mecanismos de reparación que revisan y corrigen la secuencia de DNA copiada. Los errores espontáneos en la replicación se producen generalmente por la presencia de formas tautoméricas de las bases nucleotídicas que son capaces de aparearse con bases que no corresponden. Estas lesiones al DNA pueden estar favorecidas por mutágenos endógenos, tales como radicales libres, o exógenos, tales como análogos de bases y agentes intercalantes, que

generan cambios en las propiedades químicas de las bases (revisión completa de las mutaciones espóntaneas en Maki (2002)).

Debido al carácter degenerado del código genético y de la presencia de regiones regulatorias, codificantes y no codificantes existe una clasificación de las mutaciones en términos de la función alterada (Griffiths y col., 2005): A) Mutaciones silentes que modifican la secuencia de DNA codificante en una base que no cambia el aminoácido codificado por ese codón; o mutaciones en regiones regulatorias que no alteran la expresión de ningún gen. B) Mutaciones sin sentido que alteran un triplete generando un codón de término dejando una proteína incompleta. C) Mutaciones de cambio de sentido que cambian el aminoácido codificado generando una proteína no funcional o con su función modificada. D) Mutaciones de cambio de fase producidas por la inserción o deleción de cualquier número de bases que no sea múltiplo de tres que generan una modificación de la secuencia aminoacídica a partir de la inserción (o deleción) resultante.

Además, estas mutaciones se pueden clasificar en sustituciones sinónimas y no sinónimas. El primer tipo ocurre cuando el aminoácido es el mismo o el nuevo tiene propiedades similares que no afectan la función de la proteína. En las sustituciones no sinónimas, el aminoácido modificado es químicamente distinto y puede producir un cambio severo en la estructura y función de la proteína.

Duplicaciones

Existen diversos mecanismos duplicativos descritos (revisión completa en Hastings y col. (2009)). Las duplicaciones en tandem se refieren a la presencia de múltiples copias de un segmento de DNA adyacentes entre sí. Se originan principalmente durante la recombinación de hebras, donde una de ellas presenta repeticiones cortas que generan un cross over desigual y por lo tanto una hebra queda con una copia extra (fig. 1.2).

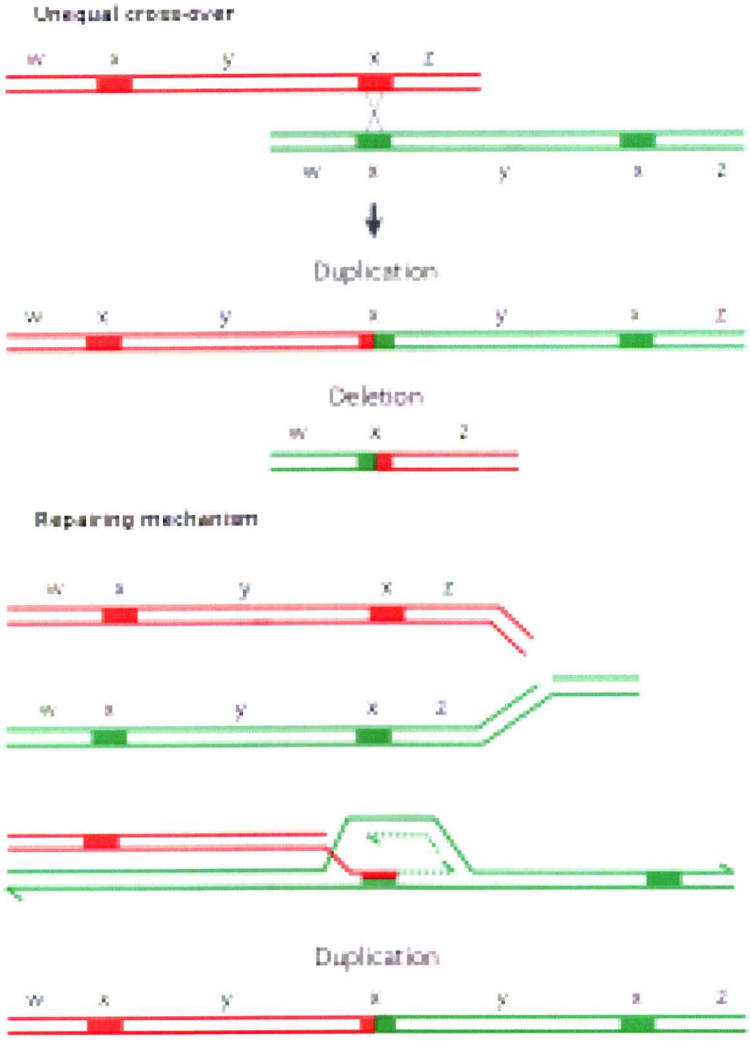

Figura 1.2: **Cambios en el número de copias por recombinación homóloga.** Las líneas representan una hebra de DNA en las cuales el extremo 3' esta indicado por la punta de la flecha. Se observa como ocurre NAHR por crossover desigual donde la presencia de un segmento repetido resulta en productos que estan recíprocamente duplicados y delecionados para esta secuencia. En la figura de abajo se muestra como ocurre la duplicación al momento de reparar homólogamente una hebra de DNA dañada. Figura modificada de Hastings y col. (2009).

Otro mecanismo de duplicación parcial es la retrotransposición. Este evento ocurre cuando un mRNA es retrotranscrito a DNA complementario (cDNA) y luego se inserta al azar en cualquier lugar del genoma. A nivel de secuencia, este mecanismo se identifica por el hecho que la nueva copia no presenta intrones, ni regiones regulatorias, por lo que es improbable que esta copia pueda ser transcrita transformándose casi inmediatamente en un **pseudogen** (Zhang, 2003).

El evento de duplicación de un genoma completo o WGD mantiene la razón entre todos sus genes, ya que el número de todos ellos se ha doblado por igual. Los mecanismos por el que ocurre se denominan alopoliploidía y autopoliploidía. La alopoliploidía se produce como consecuencia de hibridación interespecífica . La autopoliploidía se produce cuando un set de cromosomas en una especie se duplica normalmente durante el ciclo celular, y luego falla a nivel de meiosis, generando gametos diploides que al fusionarse forman un tetraploide. Si falla la mitosis, la célula no se divide correctamente manteniendo en duplicado el set cromosómico del organismo (Griffiths y col., 2005).

Se ha propuesto que luego de un evento duplicativo debe ocurrir la rediploidización del genoma debido a la naturaleza de la mitosis y al apareamiento par específico de cromosomas homólogos durante la meiosis. El mecanismo por el que ocurre esto no esta completamente establecido, se cree que son múltiples procesos los que actúan y generan la pérdida de secuencias duplicadas (pseudogenización). Un buen ejemplo viene de los salmones, los cuales a pesar de presentar una gran diversidad en su número diploide 2n, han mantenido un número constante de brazos cromosómicos (NF = 100).

1.3.2. Consecuencias y preservación de duplicados

Las consecuencias generadas por un evento duplicativo dependen principalmente de la naturaleza de éste. Las duplicaciones en tandem cambian el número de copias local de algún gen produciendo un desbalance en la proporción génica que puede ser perjudicial para la célula. Este tipo de duplicaciones favorece la recombinación desigual durante el crossover aumentando acumulativamente el número de duplicados (Ohno, 1970), o generando una predisposición a producir duplicaciones mayores (ver fig. 1.2). Se ha propuesto que la duplicación de exones en tandem ha dado origen a algunas formas de splicing alternativo (Kondrashov & Koonin, 2001), por lo que estos eventos están sujetos a su efecto fenotípico inmediato bajo la presión selectiva actuante y no necesariamente son nocivos. Su persistencia en el genoma de una especie está dado por la probabilidad de **fijación**. El tiempo en que esto ocurra estará sujeto a la **deriva génica** y la selección natural (Innan & Kondrashov, 2010).

El aumento en el número del set completo de cromosomas, WGD, y sus consecuencias se ha estudiado ampliamente en los últimos años (revisión completa en Semon & Wolfe (2007)). Se ha observado que los primeros cambios luego de este evento son rearreglos cromosómicos. Se estima que la persistencia de genes duplicados generados por este medio es mayor en comparación con las duplicaciones parciales ya que se mantiene la relación en el número de genes y se mantienen las regiones regulatorias y no codificantes asociadas a los genes duplicados permitiendo así su expresión. Además, provee material crudo de secuencia donde innovar producto de la relajación en la presión selectiva actuante (ya existe un locus funcional) (Chain y col., 2008).

Otro evento evolutivo relacionado con WGD es la especiación. Los peces teleósteos representan en términos de número de especies el grupo más exitoso de los vertebrados con aproximadamente 25 000 representantes con una gran diversidad morfológica entre

sí (Volff, 2005). Una relación propuesta entre esto y los eventos de duplicación es que, una vez que el organismo presente dos copias de un mismo gen, sólo una de ellas se perderá y la otra copia continuará cumpliendo la función. Cada población perderá una de las copias arbitrariamente dependiendo de las mutaciones que cada una de ellas haya acumulado y la ventaja selectiva que le confieran sobre la otra. Esto dependerá del medio ambiente en que la población se desenvuelva por lo que cada una puede perder una copia distinta con una función ligeramente modificada sobre la ancestral, este proceso se ha denominado Reciprocal Gene Loss (RGL) (Semon & Wolfe, 2007).

Los genes duplicados que se retengan acumularán distintas mutaciones, y por ende, su función se modificará diferencialmente. Esto representa una fuente de divergencia donde algunas poblaciones presentarán variantes distintas, ya sea a nivel codificante o regulatorio. Existen varios modelos descritos que explican la preservación de genes duplicados (revisados por Conant & Wolfe (2008)).

El destino más común para una gen duplicado es la pseudogenización, es decir, acumular mutaciones deletéreas que lo convierten en un parálogo no funcional (modelo clásico de no-funcionalización propuesto por Ohno (1970)). A partir de esto, se esperaría que la tasa de retención de duplicados en todas las especies fuese baja, lo que en la realidad no se observa. Se estima que aproximadamente la mitad de los genes vertebrados se han mantenido como duplicados funcionales (Nadeau & Sankoff, 1997). Se ha demostrado que un 15% de los genes humanos son duplicados, en *D. rerio* al menos un 20%, en *Arabidopsis* un 30% (revisado en Prince & Pickett (2002)) y en salmones un 50% (Utter y col., 1973).

Para estudiar el destino de un par de genes duplicados es necesario considerar tres períodos de vida del par: creación, preservación-fijación y optimización (Hastings y col., 2009). La creación del par está sujeta al mecanismo de duplicación por el que pasa el

organismo. En la fase de preservación-fijación se previene la pérdida neutral de uno de los duplicados y se evita la **selección purificadora** que eliminaría las mutaciones que se presenten en la población y modifiquen el gen. Gracias a que existen dos copias funcionales (con una función distinta o con una de sus funciones optimizadas dado por nuevas mutaciones) se mantienen ambas. La optimización se refiere al proceso opcional en el que una de las copias continúa evolucionando en favor de alguna función especializada, es decir, puede acumular más mutaciones que son fijadas en la población a través de la **selección positiva**.

El primer modelo descrito se denominó neofuncionalización (Ohno, 1970), en el que una vez ocurre la duplicación, una de las copias adquiere mutaciones que le confieren una nueva función. Si esta mutación le confiere una ventaja adaptativa inmediata será fijada por la selección positiva, si no es una ventaja inmediata estará sujeta a la **deriva génica**. En este caso, Conant & Wagner (2003) discuten la aleatoriedad del evento duplicativo argumentando que si bien las mutaciones son al azar, la nueva función requerida está relacionada con las funciones del gen ancestral que se duplica.

Entre los modelos actuales se encuentra el de subfuncionalización por pérdida de función, o **DDC** (Duplication, Degeneration, Complementation Process), y está basado en la adquisición de mutaciones degenerativas neutrales en múltiples regiones regulatorias que cumplen subfunciones complementarias, lo que genera una presión selectiva que obliga a la preservación de ambas copias. Una de ellas realiza una función distinta a la otra, por lo que ambas son necesarias para cumplir la función del gen (Force y col., 1999). El modelo de subfuncionalización por mutaciones adaptativas (Des Marais & Rausher, 2008), también denominado escape del conflicto adaptativo (**EAC**, escape from adaptive conflict) se refiere al proceso en el cual, el gen previo a la duplicación codifica para una proteína que cumple múltiples funciones, las mutaciones que ocurren luego del evento

optimizan diferentes funciones en cada copia, siendo luego seleccionadas y fijadas en la población por selección direccional.

Varios autores indican que es difícil determinar cuál es el modelo de preservación por el que se han mantenido ambos parálogos, En cada modelo puede existir una etapa de neofuncionalización que encubre la función original por la que fue preservada cada copia. Para generalizar esta ambiguedad se ha propuesto el término subneofuncionalización (He & Zhang, 2005; Rastogi & Liberles, 2005), donde se combinan los modelos de subfuncionalización planteados con la adquisición previa, o posterior, de nuevas mutaciones que confieran ventajas selectivas directas a cada copia.

1.3.3. Mutaciones y selección

Actualmente, existen dos corrientes principales respecto a las teorías de evolución molecular, la primera se denomina la teoría neutral, que indica que la mayoría de la variación en una especie o entre especies es neutral, es decir, las mutaciones no modifican la capacidad de supervivencia del individuo (Kimura, 1983). La segunda plantea que la mayor proporción de la variación observada si afecta la capacidad de supervivencia de los individuos de la especie y está sujeta a las fuerzas de selección (Gillespie, 1994).

En términos evolutivos, las mutaciones representan el motor de cambio y de adaptación de nuevas funciones en una población determinada. Su preservación, fijación y localización a lo largo de una secuencia está fuertemente determinada por el mecanismo y por las fuerzas de selección actuando sobre ésta. Una mutación que confiere una ventaja adaptativa será mantenida y fijada en la población por la selección positiva. Una mutación deletérea estará sujeta a las fuerzas de la **selección negativa**. La selección purificadora es indistinta de esta última y actúa en función de eliminar los

alelos no ventajosos para la población. La discusión se centra en que proporción de estas mutaciones son realmente neutrales y cuáles no.

El test más usado actualmente para detectar la selección está basado en genómica comparativa y no necesita datos poblacionales. Compara la razón entre el número de mutaciones no sinónimas por sitio no sinónimo (**Ka**) y el número **Ks** de mutaciones sinónimas por sitio sinónimo en una secuencia codificante dada (la razón se formula dN/dS o Ka/Ks) (Yang & Bielawski, 2000). Este valor se utiliza para estimar la presencia de la selección, bajo el modelo más simple la razón $dN/dS = 1$ indica que la secuencia evoluciona bajo el modelo neutral donde no existe presión selectiva actuando. Si la razón es mayor a uno, indica que se han producido más mutaciones no sinónimas que han alterado la función de la proteína y que han sido seleccionadas positivamente. Si es menor a uno, es porque existe una selección negativa que elimina las mutaciones deletéreas de la población permitiendo sólo mutaciones sinónimas que no afectan la proteína.

La mayoría de las proteínas funcionales se verán afectadas inmediatamente una vez que adquieran una mutación no sinónima, por lo que la selección negativa domina el camino evolutivo de cada secuencia y el valor calculado de Ka/Ks será casi siempre menor a uno, descartando la selección positiva que sí puede estar actuando en otros sitios particulares del gen. Para corregir este problema se han desarrollado algunos métodos que incorporan la distribución del valor de Ka/Ks en cada sitio polimórfico por cada codón interrogado (Nielsen, 2005; Yang & Bielawski, 2000).

Actualmente, existen múltiples bases de datos que recopilan la información proveniente de las mutaciones encontradas en distintas poblaciones de distintas especies que dan cuenta de la variabilidad presente. De esta forma, los SNPs se han transformado en marcadores moleculares que, entre otras cosas, se pueden asociar a enfermedades o

a rasgos fenotípicos de importancia. Las fuerzas selectivas actúan sobre estos cambios disminuyendo la frecuencia de los alelos que reducen la capacidad adaptativa del individuo. Se ha propuesto que la búsqueda de SNPs y la detección de fuerzas selectivas en genes candidatos permitirá identificar las causas y la predisposición de cada individuo para ciertas enfermedades genéticas (Nielsen y col., 2007). Además, los SNPs capturan la información asociada a la historia y origen de la estructura poblacional de una especie determinada permitiendo realizar trazabilidad genética o estudios de pedigree que permitan la identificación de los individuos presentes en ella (Vignal y col., 2002).

1.3.4. Variación intragenómica individual

La mayoría de los estudios actuales se centran en comparar la distribución de las frecuencias alélicas de SNPs, entre poblaciones de una especie determinada o entre distintas especies, denominado variación intraespecífica e interespecífica, respectivamente (Nielsen, 2005). Ambos términos están basados en la comparación del mismo gen proveniente de muchos individuos o, entre genes ortólogos de distintas especies con genomas distintos, este tipo de variación se puede clasificar como intergenómica.

La presencia de genes parálogos que pueden presentar desde dos a múltiples copias en el genoma de un sólo individuo y las cuales, conservan algún grado de relación entre sí (función modificada, subfunción ancestral, expresión tejido específica, etc.) permite plantear el concepto de variación intragénomica. Los modelos de preservación de genes duplicados plantean la necesidad de adquirir mutaciones que les permitan evitar la selección purificadora. Estas mutaciones, con el paso del tiempo, se reflejan como SNPs (una vez alcancen un **MAF** $> 1\,\%$) y representan la variabilidad adquirida por estas copias. Además, estos polimorfismos se pueden clasificar dependiendo de su origen.

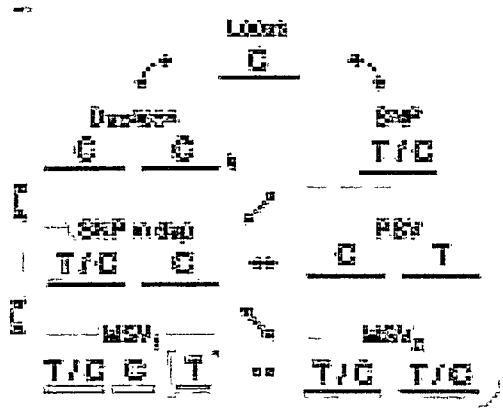

Figura 1.3: **Origen de las variantes intragenómicas.** La figura muestra el efecto que tiene la duplicación sobre los sitios que contienen SNPs, cuando una región con algún sitio polimórfico se duplica, el polimorfismo se puede encontrar en ambos paralogos o sólo uno(MSVs), en cambio, cuando uno de los duplicados diverge la diferencia se observará como un PSV. Figura modificada de Fredman y col. (2004).

Dentro de las variantes intergenómicas se encuentran dos grandes categorías: A) SNPs que producen diferencias en la expresión o función de la misma proteína en dos individuos distintos (o dos especies distintas), siendo este tipo el de mayor interés productivo ya que permite relacionar las causas de las diferencias fenotípicas en rasgos tales como crecimiento, desarrollo, color de la carne, entre otros, a un genotipo determinado. B) Variantes que no afectan el fenotipo de la proteína por ser mutaciones sinónimas que, debido al carácter degenerado del código genético, no varían el aminoácido codificado o también, debido a que éstas caen sobre una región no codificante o no regulatoria (Vignal y col., 2002). Estos SNPs son útiles como marcadores ya que pueden estar ligados a otros determinantes genéticos que modifican el fenotipo.

Dentro de las variantes intragenómicas existen (ver fig. 1.3): A) SNPs que provienen de un locus polimórfico al momento del evento duplicativo y mantienen este carácter en ambas copias, se denominan **MSVs** (Multiple Sequence Variants). B) SNPs que se encuentran debido a diferencias en una base originadas por la presencia de copias

parálogas que se han preservado en el genoma y se denominan **PSVs** (Paralogous Sequence Variants). C) Polimorfismos que se observan por la presencia de dos alelos distintos donde el individuo es heterocigoto para ese locus y corresponden a los SNPs comunes, estos últimos provienen de la variación intergenómica donde un alelo viene del padre y el otro de la madre (Moen y col., 2008).

1.3.5. Detección y predicción

Detección de regiones parálogas

La detección o búsqueda de segmentos duplicados utilizando herramientas *in silico* se ha basado en determinar homología entre secuencias de familias génicas en distintas especies para luego construir árboles filogenéticos que permitan inferir el tiempo del evento duplicativo (Robinson-Rechavi y col., 2004). Existen múltiples estudios que utilizan esta estrategia en diversas especies tales como *A. thaliana*, *S. cerevisiae* (revisado por Wolfe & Seoighe (1999)), *D. melanogaster* (Conant & Wagner, 2003). Se basan en el alineamiento de regiones cromosómicas distintas con similitud de genes que se encuentren idealmente en el mismo sentido.

Para conocer si las duplicaciones ocurrieron en un mismo período evolutivo (una sola WGD), o si provienen de múltiples duplicaciones ocurridas en tiempos distintos, Vandepoele y col. (2004) construyeron árboles filogenéticos para todas las familias génicas del genoma de Fugu que presentan entre 2 y 10 duplicados y estimaron el tiempo de ocurrencia de esta separación. Para calibrar sus árboles utilizaron la fecha de separación entre vertebrados terrestres y peces (450 MYA), y concluyeron que la mayoría de los eventos duplicativos ocurrieron hace 320MYA, tiempo estimado de la ronda duplicativa específica del linaje de los teleósteos.

La caracterización de mapas físicos también ha permitido determinar el nivel de

duplicaciones presentes en un organismo. Palti y col. (2004) utilizaron una librería BAC de trucha arcoiris para hibridar sondas específicas con los clones de la librería. Luego de generar el ensamblaje correspondiente de cada clon, encontraron que dos tercios de los loci investigados caían en múltiples **contigs** sugiriendo que estos provenían de regiones duplicadas.

Búsqueda de SNPs

La principal estrategia a la hora de buscar posibles SNPs es a partir del alineamiento de un alto número de secuencias que representen una misma región genómica. Para esto se han alineado y analizado transcritos de ESTs, los cuales provienen de secuencias expresadas de un mismo gen en múltiples individuos (Buetow y col., 1999; Zhao y col., 2003; Hayes y col., 2007b). Es posible utilizar también clones BAC, **YAC** o plásmidos y alinear las secuencias clonadas, lo que permite encontrar SNPs en regiones no codificantes.

Específicamente, al realizar un alineamiento múltiple es posible observar diferencias en ciertas bases nucleotídicas respecto del consenso que pueden considerarse *a priori* como SNPs. Para acertar en la predicción es necesario establecer parámetros que permitan diferenciar si el origen de este SNP es un artefacto propio del proceso de secuenciación o tiene un origen biológico. Este origen está dado por las variantes generadas por una mutación fijada en genes no duplicados o por polimorfismos en una región duplicada que debido a presiones selectivas mantuvo la mutación en una de las copias y se preservó a lo largo del tiempo (Ryynanen & Primmer, 2006).

Una vez realizado el alineamiento y ensamblaje de secuencias de DNA con programas especializados como CAP3 (Huang & Madan, 1999) o phrap (Gordon y col., 1999), es necesario buscar los sitios polimórficos con algún método bioinformático.

Existen programas como PolyPhred (Nickerson y col., 1997) o PolyBayes que han sido ampliamente usados en la búsqueda de SNPs.

Una vez identificados los SNPs putativos, es necesario realizar la validación experimental en la población. Esto se denomina genotipado y permite saber qué alelos lleva el individuo por cada SNP interrogado. Una vez amplificada la región candidata, existen varias técnicas de validación o genotipado. Estas incluyen electrofororesis en gel de gradiente denaturante, corte enzimático o químico, análisis de heteroduplex, análisis de conformaciones de DNA, hibridación a oligos y secuenciación del DNA (Nickerson y col., 1997). Cada una de estas estrategias presenta ventajas y desventajas. La secuenciación permite identificar el contexto genómico presente alrededor del SNP, permite observar si existen más posiciones polimórficas y su aplicación es relativamente fácil pero costosa si se quiere realizar a gran escala. Diversos métodos de genotipado se discuten en Kim & Misra (2007).

La búsqueda de SNPs ha estado enfocada principalmente en encontrar la variación entre individuos distintos ya sea de la misma especie o no. Esto ha provocado que los SNPs presentes en regiones duplicadas (PSVs o MSVs) sean confundidos con SNPs comunes. Esto conduce a errores al momento de genotipar individuos ya que en la mayoría de las técnicas usadas sólo se interroga por la presencia o ausencia de alelos. Para un PSV el individuo siempre será heterocigoto debido a la presencia de duplicados, para un MSV es más complejo debido a que puede presentar todos los genotipos, pero generalmente está ligado a una mayor presencia de heterocigotos. Tal asignación genera errores al momento de realizar estudios de mapeo, de trazabilidad o de asignación de parentesco por lo que estos SNPs deben ser evitados al momento de genotipar. Se ha estimado que este tipo de polimorfismos representan alrededor del 25% de los SNPs depositados en las bases de datos públicas (Gut & Lathrop, 2004).

1.4. Búsqueda y uso de marcadores en salmón

Una de la principales directrices en las líneas de investigación actuales sobre salmónidos es la búsqueda de marcadores moleculares que permitan realizar trazabilidad genética individual y selección genética asistida por marcadores asociados a rasgos de importancia económica (Hastein y col., 2001). Ambas líneas buscan impulsar los programas de mejora genética utilizando la genética clásica con herramientas de genética molecular.

En la última década los microsatélites, o SSRs (Short Sequence Repeats), han sido usados ampliamente como marcadores moleculares. Estos presentan una secuencia que va desde 1 a 6 pares de bases repetidas N veces, con un N variable entre individuos no emparentados. Esto ha permitido realizar análisis de parentezco y paternidad ya que se encuentran distribuidos en los genomas de todas las especies como una forma de elementos repetitivos.

Los SNPs también han sido utilizados para realizar trazabilidad en diversas especies (Goffaux y col., 2005; Vignal y col., 2002), incluyendo peces (Hayes y col., 2005, 2006). En salmón estos marcadores han permitido realizar estudios de asociación para rasgos como tasa de crecimiento, patrón de pigmentación, coloración, entre otros (Boulding y col., 2008).

Sin embargo, en salmones se ha propuesto que una combinación de SNPs y SSRs y el uso de un alto número de loci entrega mejores resultados que el uso de ambos tipos de marcadores por sí solos (Ryynanen y col., 2007; Grandjean y col., 2009; Rengmark, 2006). La principal diferencia radica en que los microsatélites son altamente polimórficos y se requiere un número menor de estos en comparación con los SNPs (Tsuchihashi & Dracopoli, 2002), pero su costo de genotipado por unidad es mayor.

En los últimos años numerosos estudios se han dedicado a la búsqueda y validación

de SNPs en salmón. Hayes y col. (2007a) publicaron 2507 posibles SNPs a partir del alineamiento de secuencias expresadas (Expressed sequence tags, ESTs) del salmón del Atlántico. De estos SNPs, ochenta y seis fueron seleccionados para ser validados, de los cuales el 74% resultaron positivos y de éstos, un 14% resultó con exceso de heterocigocidad lo que sugiere que las regiones que albergan estos SNPs se encuentran duplicadas (PSV o MSV).

La presencia de gran cantidad de elementos transponibles y repetitivos a lo largo del DNA del salmón, así como también la existencia de extensas regiones parálogas, plantea un desafío importante a la hora de diferenciar el origen de un posible SNP. Como se ha descrito, el carácter polimórfico de estos elementos genera diferencias en las secuencias debido a que un mismo elemento se puede encontrar repetido múltiples veces en el genoma de un mismo individuo y así dar indicios de SNPs "falsos"(PSVs), sólo por efecto de la redundancia del alineamiento realizado con aquellas secuencias (Gut & Lathrop, 2004; Ryynanen & Primmer, 2006; Smith, 2005; Smith y col., 2005).

1.5. Objetivos e hipótesis

El análisis dado por la estrategia bioinformática utilizada junto con las evidencias presentadas permitirá la predicción de loci parálogos basándose en la siguiente hipótesis: **Los SNPs presentes en el genoma de un único individuo provienen mayoritariamente de la variación intragenómica dada por las regiones parálogas que han escapado a la selección purificadora y se han preservado a lo largo de la evolución.**

El objetivo general de la investigación es predecir regiones parálogas a través del ensamblaje de secuencias genómicas de un único salmón en combinación con la predicción de SNPs, filtrado de elementos repetitivos y anotación genómica (fig. 2.1).

Los objetivos específicos son: 1) Generar un ensamblaje de secuencias genómicas del salmón del Atlántico. 2) Establecer una distribución de SNPs en secuencias genómicas proveniente de un único individuo. 3) Observar la distribución de elementos repetitivos y elementos transponibles. 4) Anotar genes conocidos por búsqueda de similitud con blast. 5) Establecer un patrón conservado de polimorfismos que permita predecir loci parálogos 6) Evaluar los modelos de preservación de genes en regiones duplicadas. 7) Determinar las mejoras metodológicas aplicables a la búsqueda *de novo* de SNPs. 8) Determinar la presencia de fuerzas selectivas actuando sobre duplicados (Razón Ka/Ks).

2. Materiales y métodos

2.1. Análisis bioinformático

2.1.1. Alineamiento y ensamblaje

Se obtuvieron las 217632 secuencias de los extremos de clones de la librería **BAC**, denominadas de ahora en adelante **BES** (Bac End Sequences), proveniente de un trabajo previo realizado por Thorsen y col. (2005). Los insertos clonados tienen un tamaño promedio de 180 Kb, los extremos fueron secuenciados y se obtuvo un largo promedio de 900 pb. La librería BAC presenta una cobertura de 18X y los BES fueron solicitados via ftp al laboratorio a cargo del Dr. Willie Davidson en Simon Fraser University, Canadá. Los 217632 **cromatogramas** BES fueron evaluados usando la estrategia presentada en la figura 2.1, donde el primer paso es realizado con el programa phred (Ewing y col., 1998) que permite la asignación de un valor de calidad, con rango de 4 a 60, a una base nucleotídica proveniente de una máquina de secuenciación. El valor de calidad Q es la probabilidad de error transformada al logaritmo en base 10 (Un valor Q de 20 indica que existe una certeza del 99 % que la base indicada sea correcta, un valor Q de 30 indicaría un 99,9 % y así sucesivamente). Luego de asignar una medida de calidad, es necesario **enmascarar** el vector pTARBAC, utilizado en la construcción de la librería, con el programa Crossmatch para luego realizar el alineamiento final y ensamblaje en **contigs**

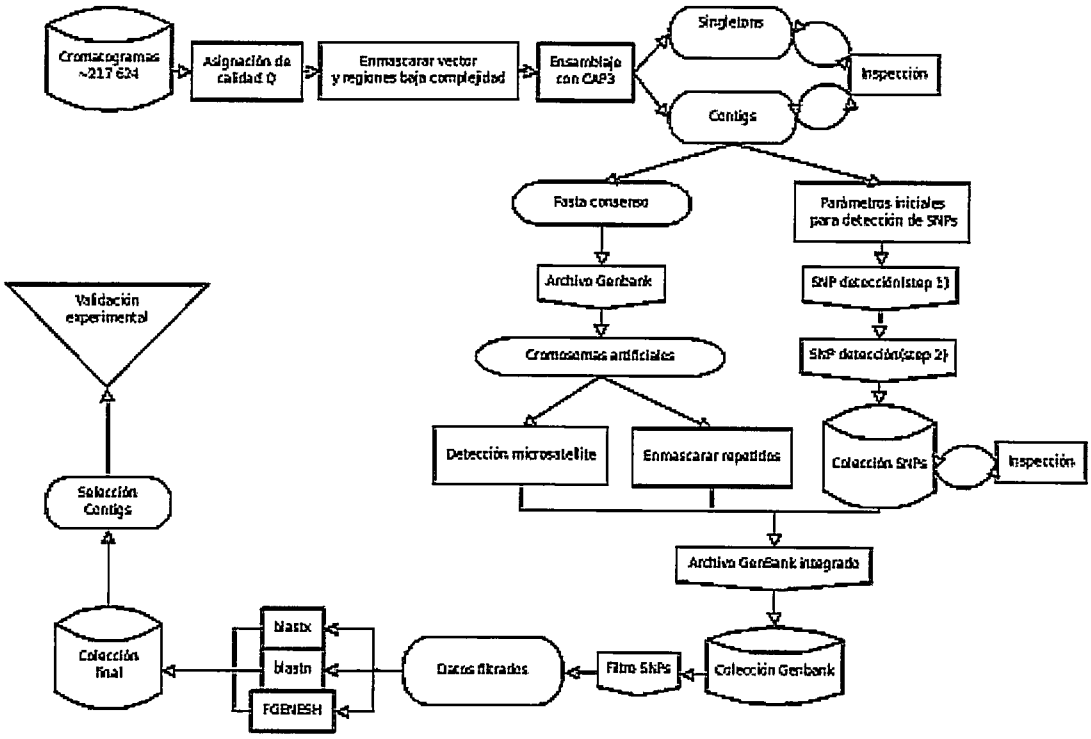

Figura 2.1: **Pipeline ParalogIt**. Estrategia bioinformática para búsqueda de SNPs y predicción de regiones duplicadas.

con el programa CAP3 (Huang & Madan, 1999). El parámetro de sobrelape entre lecturas fue de 130pb con un porcentaje de identidad del 95%. El ensamblaje fue visualizado y depurado con el programa Consed (Gordon y col., 1999). La estrategia implementada se observa en la figura 2.1.

2.1.2. Elementos repetitivos

Las **secuencias consenso** de los contigs generados fueron inspeccionadas para establecer el contenido de elementos repetitivos presentes en el genoma del salmón usando para esto la herramienta bioinformática RepeatMasker (Smit y col., 1996). Los tipos de elementos repetitivos encontrados por este programa se observan en la sección resultados en la figura 3.2 y se pueden agrupar en 2 categorías, transposones y no transposones. Para el análisis se utilizaron 3 bases de datos: Zebrafish (*Danio rerio*) y Fugu (*T. rubripes*) provenientes de la Repbase v.20080801, y una base de datos desarrollada por cGRASP específica para repeticiones en salmónidos.

Se utilizó además el programa Tandem Repeat Finder (Benson, 1999) para identificar repeticiones en **tandem**, tales como microsatélites, para delimitarlos, establecer el largo del patrón y el número de veces que éste se repite (número de copias). De esta forma, se desarrolló una base de datos con la información obtenida en la búsqueda de elementos repetidos para luego integrarla con los datos obtenidos en la búsqueda de SNPs (Colección Genbank en la figura 2.1).

2.1.3. Búsqueda de SNPs

Para poder establecer la presencia de polimorfismos en el ensamblaje obtenido se utilizó el script AnalyzeSNPs del paquete toAmos (Pop y col., 2004) y se filtraron los resultados usando un script *in-house* **BioPerl** (Fase 1 y 2 de detección de SNPs en la figura 2.1), bajo los siguientes parámetros: calidad mínima Q promedio de 20 entre todas las bases ubicadas en la misma posición; nivel de **cobertura** mínimo de 4 del contig; y **porcentaje de variación** esperado de al menos un 20 %, esto corresponde a la fase 1 de la detección de SNPs. En la fase 2, los SNPs fueron clasificados en tres categorías para observar el efecto que tienen estos parámetros en la predicción: A) Aquellos que

presenten una calidad Q igual o mayor a 40 con cobertura mayor o igual a 4; B) SNPs con una calidad Q mayor a 30 y con cobertura mínima de 6 y C) SNPs con calidad Q mayor a 20 pero con cobertura mínima de 8. Todos los SNPs encontrados fueron inspeccionados manualmente, utilizando el programa Consed, para evitar lecturas falsas o incorrectas, para luego ser integrados a la colección Genbank mencionada.

2.1.4. Anotación de genes, regiones codificantes y elementos repetitivos.

El filtrado de la colección GenBank se realizó mediante la implementación de un script (filtro SNP en la figura 2.1) que permite seleccionar las secuencias que cumplan con ciertos parámetros establecidos dados por la información presente (La colección GenBank contiene información de secuencias nucleotídicas, SNPs y elementos repetidos). El filtrado final considera tomar solamente los contigs que presenten al menos un SNP de cualquiera de las categorías, y a su vez, que este SNP no esté dentro de la secuencia de un elemento repetitivo.

Para la anotación de genes se utilizaron variantes del programa Blast (McGinnis & Madden, 2004) (fig. 2.1). BlastX requiere secuencias nucleotídicas de entrada, las traduce a proteínas, y las compara con secuencias de proteínas de distintas bases de datos escogidas. De esta forma, cada contig fue tabulado con su número de SNPs, hit blast en contra de las bases de datos nr, Swissprott y EC. Para los contigs que no presentaron hit se realizó una búsqueda blastn en contra de las secuencias de núcleotidos presentes en NCBI.

Se utilizó el programa FGENESH para la predicción de **CDS** en las secuencias de DNA de cada contig. Los resultados fueron tabulados manualmente para seleccionar el mejor hit blast con e-value $< 1e^{-}5$.

El tipo de secuencia que alberga el SNP se clasificó como: I) Desconocido, para

contigs que no presentan CDS o hit blast cercano al SNP. II) Repetitivo, para SNPs encontrados en transposasas, microsatélites u otros elementos repetidos que no fueron filtrados previamente. III) Intrón, para SNPs que se ubican entre dos CDS o están cerca de un gen y su secuencia no corresponde a un exón terminal o inicial de la proteína. IV) Codificante, SNPs dentro de un CDS predicho o algún gen encontrado por blast. De esta forma se obtuvo la colección de datos final que incluye los contigs con toda la información encontrada.

2.1.5. Visualización y selección de contigs para validación experimental de SNPs

La visualización de la colección de contigs se realizó con el programa ARGO (Engels y col., 2006) desarrollado por el MIT. A partir de la colección de datos final (fig. 2.1) se seleccionaron 10 contigs (Cuadro 3.1) para realizar el genotipado de los SNPs, basándose en los siguientes criterios: Presencia de uno o más SNPs dados por las tres categorías establecidas previamente; ausencia de elementos repetidos que no hubiesen sido filtrados previamente tales como transposasas u otros, posibilidad de diseñar partidores específicos para realizar la amplificación por PCR del locus candidato y que estos contigs presenten algún gen o CDS anotado cercano al SNP. Para simplifcar el análisis, estos contigs fueron numerados del 1 al 10 como se observa en el cuadro 3.1.

Para determinar qué exón estaba presente en cada contig seleccionado se obtuvo la secuencia aminoacídica de la proteína homóloga de *D. rerio* y la secuencia génomica que alberga el gen. Ambas secuencias se alinearon con el programa WISE2 (Birney & Copley, 2001) que entrega todos los posibles exones e intrones del gen. De esta forma se corrigió manualmente el alineamiento obtenido inicialmente entre el contig y su hit blast, y se determinó la posición de inicio y final del exón presente a lo largo del contig. La función

y otras propiedades de la proteína se obtuvieron a partir de la información disponible en UniProt y GenBank.

Además, se obtuvieron secuencias ortólogas, a partir de NCBI, del gen *Hsp70* (Contig 6 secuenciado) de *D. rerio* y *O. mykiss*. Para esta última especie se ha descrito la presencia de un gen parálogo (Ojima y col., 2005), cuya secuencia también fue obtenida. Se descargó también una secuencia de cDNA previamente caracterizada para salmón del Atlántico y todas ellas se alinearon con el programa CLUSTALW (Thompson y col., 1994). Luego, se utilizó el programa KaKs calculator (Zhang y col., 2006) para determinar la tasa de sustituciones sinónimas por sitio sinónimo (Ks o dS) y la tasa de sustituciones no sinónimas por sitio no sinónimo (Ka o dN). La razón entre estos dos valores (Ka/Ks) se utilizó para determinar el nivel de fuerza de la presión selectiva existente en este locus. De los modelos disponibles en el programa se escogió el que considera la ocurrencia de transversiones y transiciones de manera desigual, y el sesgo en el uso de codones. Este modelo está descrito en Posada (2003).

2.2. Validación experimental

2.2.1. Diseño de partidores

Para el diseño de partidores se utilizó el programa Primer3 (Rozen & Skaletsky, 2000), los partidores fueron seleccionados basándose en el método del triplete 3′ (Yuryev, 2007), con una tm entre los 60°C y los 65°C. Además, se realizó una búsqueda blast de los posibles partidores en contra de todos los BES y secuencias del mapa físico del salmón disponible en la base de datos ASALBASE para asegurarse que la amplificación ocurra sólo en el locus objetivo. Los oligos fueron sintetizados por IBT en concentración 25nmol y fueron resuspendidos para obtener un stock de 100X, según las especificaciones del

fabricante.

2.2.2. Extracción de DNA genómico y amplificación de fragmentos candidatos

Para extraer el DNA del salmón se tomaron 3 aletas de individuos *Salmo salar* provenientes de 3 familias distintas provistas por AQUAINNOVO (Individuos A, B y C), se maceraron y luego se siguió el protocolo estándar del kit de extracción de DNA genómico DNeasy (QIAGEN). Luego se comprobó la presencia del DNA genómico en un gel de agarosa 1 %, y se cuantificó utilizando espectrofotometría (NANODROP V1000) para comprobar también, el espectro de absorción característico del DNA. El protocolo de optimización de PCR siguió tres fases. En un volumen total reacción de 25uL y con 30 ng en promedio de DNA genómico de cada individuo, se utilizó una concentración 400 nM de cada partidor, 1U de Promega Taq Polimerasa, 1,5mM de MgCl y 0,2mM de dNTPs. Las reacciones PCR fueron realizadas en termociclador PTC-2000 a intervalos de 35 segundos para el tiempo de hibridación y de extensión durante 30 ciclos (fig. 2.1). La temperatura inicial de hibridación escogida fue de 57°C para todos los partidores. Las regiones que no pudieron ser amplificadas pasaron a la segunda fase de optimización que incluyó TOUCHDOWN PCR con una temperatura de hibridación de 68°C para los primeros 10 ciclos, 63°C para los siguientes 10, y 57°C para los últimos (fig. 2.1). En la tercera fase de optimización se modificó la concentracion de MgCl a 2,5mM y de templado a 60 ng. Para resolver la presencia del producto de PCR se realizó electroforesis en gel de agarosa al 1,5 % con bromuro de etidio y se observó la presencia de las bandas bajo luz UV.

PCR cycle	Temp *C	Time
Start	95	3 min
2 - Denaturacion	95	30 seg
3 - Annealing	57	30 seg
4 - Extension	72	30 seg
5 - Go To step 2		30 times
6 - Extension	72	10 min
7 - End	15	Forever
Touchdown PCR		
Start	95	3 min
2 - Denaturacion	95	30 seg
3 - Annealing	68	35 seg
4 - Extension	72	35 seg
5 - Go To step 2		10 times
6 - Denaturacion	95	30 seg
7 - Annealing	63	35 seg
8 - Extension	72	35 seg
9 - Go To step 6		10 times
10 - Denaturacion	95	30 seg
11 - Annealing	58	35 seg
12 - Extension	72	35 seg
13 - Extension	72	10 min
14 - End	15	Forever

Cuadro 2.1: **Condiciones de PCR**

2.2.3. Purificación de DNA desde gel de agarosa

Para la purificación del producto de PCR de cada una de las regiones se extrajo cuidadosamente el trozo de gel de agarosa que contenía el producto y luego se utilizó el kit de extracción en gel QIAquick (QIAGEN), usando el protocolo estándar. El DNA fue resuspendido en un volumen final de 25 uL buffer TE y se almacenó a $-20°C$. Todas las muestras fueron cuantificadas por espectrofotometría utilizando NANODROP V1000 para asegurar la concentración mínima requerida para secuenciar.

2.2.4. Secuenciación y genotipado de SNPs

La secuenciación se realizó en ambos sentidos para cada locus y para cada individuo. Primero se secuenciaron solo las muestras con el partidor directo. Las muestras que no dieron una secuencia limpia en la region del SNP fueron purificadas nuevamente y enviadas a secuenciación con el partidor reverso junto con el resto de las muestras. Las secuencias fueron generadas con Applied Biosystems (ABI) 3730 Capillary electrophoresis Genetic Analyzer en el Departamento de Secuenciación de la Unidad de Ciencias Biológicas de la University of California, campus Davis. Los cromatogramas fueron analizados visualmente con el programa Consed para ver si era posible conocer el genotipo de cada individuo.

La secuencia final para cada individuo y para cada locus fue generada siguiendo la estrategia observada en la figura 2.2. Las secuencias de ambos sentidos fueron alineadas entre sí usando blast2seq, luego se alinearon por separado con el consenso obtenido previamente en el análisis *in silico* para así determinar la mejor secuencia final de cada individuo. Para los individuos que presentaron algún cromatograma con problemas de señal/ruido se seleccionó la secuencia de mejor calidad. Para determinar la secuencia final de cada locus se realizó un ensamblaje de todas las secuencias obtenidas utilizando Phrap. Luego, cada nueva secuencia consenso obtenida fue alineada con su respectiva secuencia de referencia para determinar su largo y la posición del SNP correspondiente. La visualización generada permite observar la presencia de heterocigotos para el SNP buscado, explorar la presencia de otros SNPs, corregir errores de lectura del cromatograma y generar la secuencia final del locus. En el caso de que un mismo locus se agrupe en contigs distintos se realinearon los consensos de ambos contigs y se generó una nueva secuencia. El genotipo se determinó basandose en la presencia de dos peaks distintos sobrelapados en la misma posición (heterocigoto).

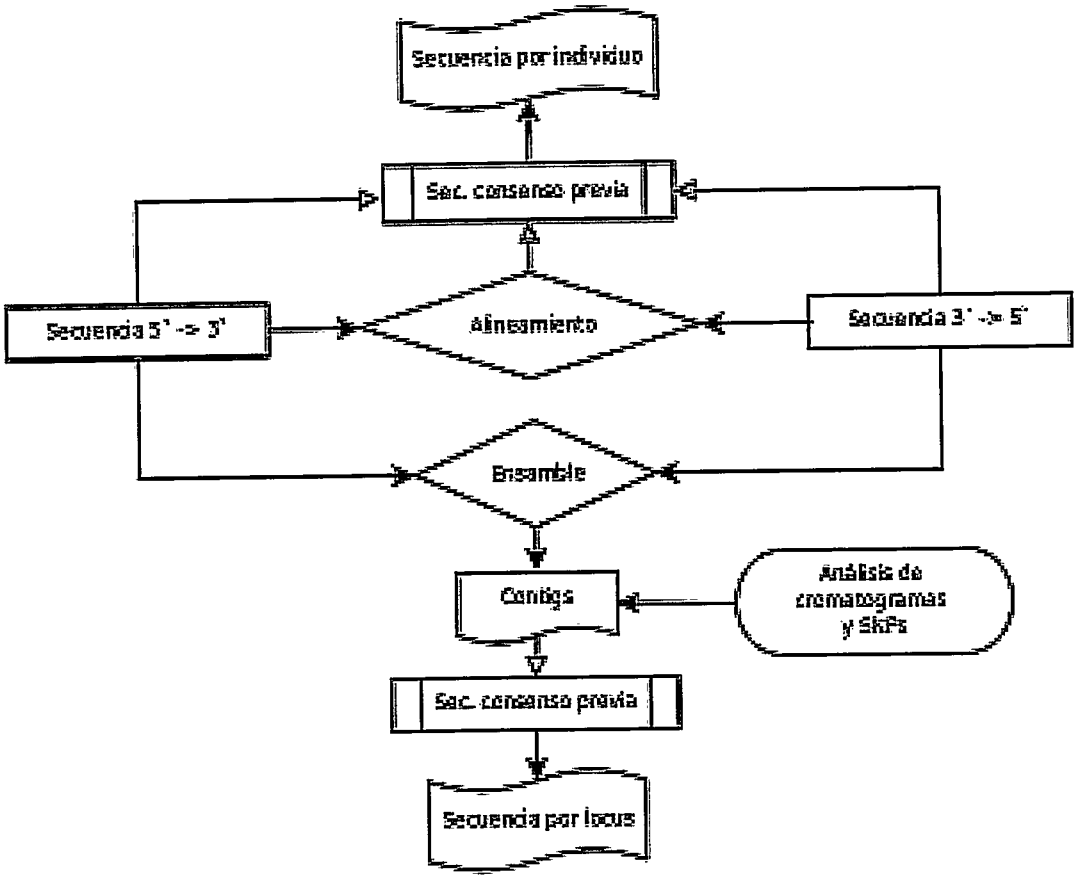

Figura 2.2: **Pipeline post-secuenciación.** Las secuencias en ambos sentidos de cada individuo son alineadas por separado para luego ser comparadas con la secuencia de referencia y así generar un consenso para el individuo. Todas las secuencias obtenidas fueron ensambladas para generar un consenso comparable a la referencia y así generar una secuencia por locus. Las flechas blancas indican alineamiento por pares.

3. Resultados y discusión

3.1. Secuencias BES

Un total de 217 624 secuencias provenientes de los extremos de los clones de una librería BAC (Thorsen y col., 2005) (cobertura $18\times$) fueron obtenidas y analizadas. El largo promedio con bases de calidad phred mayor que 20 (largo Q20) es de 669 pb, siendo un valor favorable comparado con el largo Q20 obtenido en BES de humanos (477 pb, Zhao y col. (2000)), ratón (485 pb, Zhao y col. (2001)), pez gato (557 pb, Xu y col. (2006)) y salmón del Atlántico (630 pb, Ng y col. (2005)).

El programa crossmatch filtró las regiones de baja complejidad y enmascaró el vector pTARBAC usado en la creación de la librería generando un total de secuencia genómica de alta calidad de $1{,}45 \cdot 10^8$ pb que, dada la estimación en el tamaño del genoma del salmón de $3 \cdot 10^9$ pb, representa aproximadamente un 4,9 % de éste. El porcentaje de A+T se estimó en 58,1 % y el de %G+C en 41,9 %.

3.2. Ensamblaje y visualización de las secuencias

CAP3 fue utilizado porque genera un mayor número de contigs, más cortos y secuencias consenso con menos errores en comparación con otros ensambladores. El

número de contigs obtenidos fue de 25 277 con un largo promedio de 895 pb que en total representaría un 0,75 % del genoma completo del salmón. El número de singletons (o contigs con solo una lectura) es de 124 309 con un largo promedio de 1290 pb.

Los contigs fueron inspeccionados visualmente para corroborar que no existiera alguno demasiado largo (> 3000 pb) ya que fueron generados a partir de los extremos de clones BAC que representan secuencia genómica de aproximadamente 1kb cada 20kb; solamente tres contigs no cumplieron con esta condición. El histograma de la profundidad promedio (cobertura) se observa en la figura 3.1. Ésta sólo muestra los contigs con 4 o más lecturas, el número de contigs con 2 ó 3 lecturas es de 20 066 (79 %).

Existen 115 contigs con cobertura promedio igual o mayor a 18, esto sugiere que en estos contigs se alinearon lecturas provenientes de regiones repetidas en el genoma. Esto se produce porque la enzima de restricción cortó múltiples veces una misma secuencia proveniente de loci distintos y el programa de ensamblaje sólo evalúa similitud nucleotídica y calidad juntando cada lectura en el mismo contig. Debido a que el objetivo de este trabajo es precisamente identificar regiones duplicadas, estos contigs no fueron excluidos del análisis siguiente.

3.3. Elementos repetitivos y transponibles

Los resultados del programa Repeatmasker arrojaron que al utilizar la base de datos de *T. rubripes*, un 5,05 % del total de secuencia fue identificado como repeticiones. En *D. rerio* el porcentaje es de 8,83 %. El mayor porcentaje de repeticiones encontradas con la base de datos de Zebrafish por sobre la de Fugu se puede explicar por las diferencias en el tamaño del genoma de ambas especies. El porcentaje de elementos repetidos encontrados usando las secuencias de salmónidos asciende a un 32,77 %, similar al encontrado en estudios previos donde analizan regiones genómicas de salmón (Mitchell,

2004). Este alto porcentaje sugiere una expansión de elementos repetitivos específica de salmónidos a partir del momento de divergencia entre éstos y los peces teléosteos dando cuenta de su participación en reestablecer el estado diploide luego de un evento de duplicación génomica. Estos elementos pueden participar facilitando la recombinación de regiones distintas del genoma y así estabilizar la condición dada por una meiosis aberrante producto de múltiples copias homólogas en distintos cromosomas.

La gran mayoría de estos elementos específicos de salmónidos provienen de transposones representando el 15,8 % del total de secuencia analizada (fig. 3.2). Estos elementos tienen múltiples efectos y su naturaleza es bastante diversa. Esto indica que su acumulación en el genoma de alguna especie debería estar originada por alguna propiedad que les permita preservarse y aumentar su número. Así, los elementos transponibles han cooperado con el éxito evolutivo de los salmónidos a partir del momento de divergencia con los demás vertebrados. De Boer y col. (2007) identificaron expansiones definidas de actividad transposónica que ocurrieron luego de los eventos de poliploidía y coincidieron con los eventos de especiación de los sálmonidos. Esto reafirma la hipótesis que los elementos transponibles son agentes evolutivos que modelan la arquitectura de los genomas.

Las siguientes secuencias repetitivas más representadas son retroelementos tales como SINEs, LINEs y LTRs. Sin embargo, su acumulación es similar en las tres especies indicando que estos elementos se han conservado en cantidad, y similitud de secuencia, y no son afectados por eventos duplicativos o, más bien, su rol es menor en la reestructuración de los cromosomas. También se debe considerar la cantidad de repeticiones no clasificadas presentes en el salmón que dan cuenta de un 4 % de secuencia. Estas repeticiones pueden corresponder a cualquier categoría de elementos repetitivos todavía no caracterizada que sí pueden tener algún grado de participación en

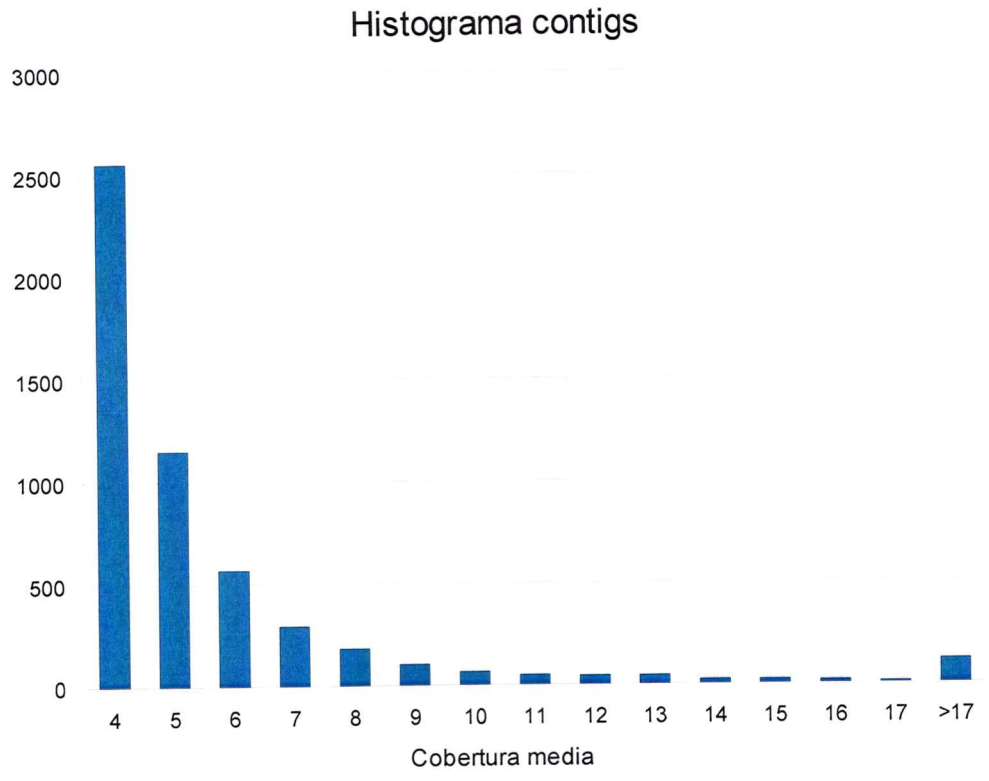

Figura 3.1: **Histograma de la profundidad promedio de contigs.** Se observa la presencia de un alto número de contigs con cobertura promedio de 4 ó 5. Se observa además que, contra lo esperado, existe un número importante de contigs que presentan una cobertura promedio mayor a 18 lecturas.

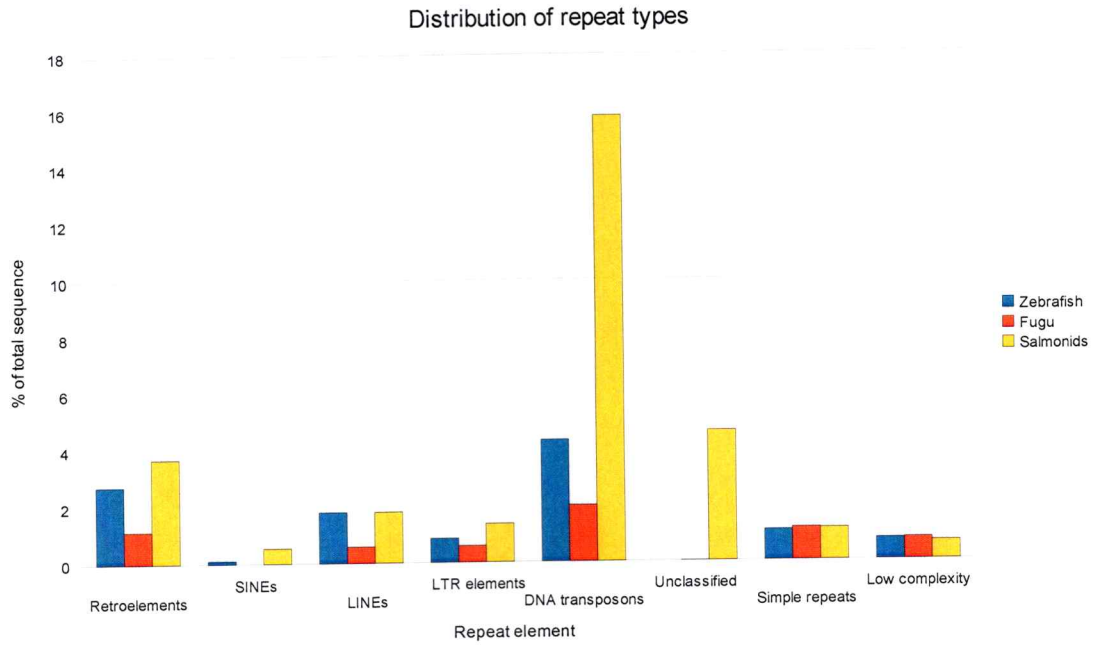

Figura 3.2: **Distribución de elementos repetitivos**. Distribucón en las secuencias genómicas del salmón según su tipo y las bases de datos disponibles, salmónidos, *D. rerio* y Fugu.

la evolución de estas especies.

Para complementar los resultados obtenidos por el programa RepeatMasker e identificar microsatélites se realizó una búsqueda *de novo* que encontró 9072 repeticiones con patrones cuyo largo varía entre 1 y 500 pb en las 25277 secuencias. La presencia de repeticiones con un patrón que va de 1 a 100 pb da cuenta del 86 % del total. El rango encontrado en el número de copias varía de 1 a 1200, sin embargo el 99,3 % de las repeticiones presenta un número de copias menor a 100. Los microsatélites (patrón de secuencia de largo 1 a 6 pb) representan un 22 % (1985), y de este, un 64 % (1266) presenta secuencias que flanquean ambos extremos necesarias para diseñar partidores. La figura 3.3 muestra la distribución de los microsátelites según el número de bases que

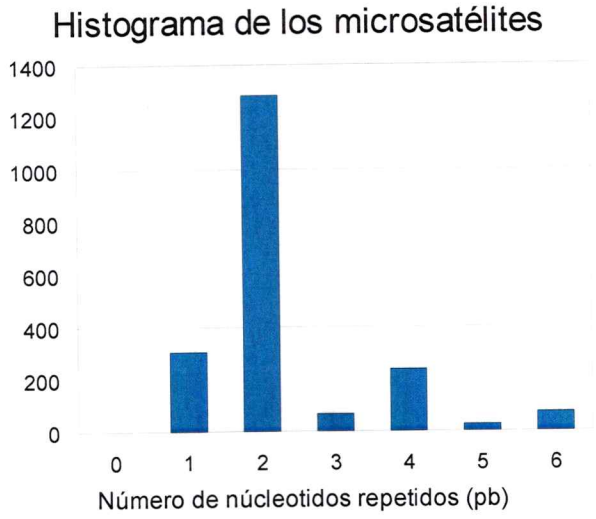

Figura 3.3: **Distribución del número de microsatélites.** Distribución en las secuencias genómicas del salmón según el patrón de bases que se repiten (programa TRF).

se repiten. El porcentaje de microsatélites encontrados respecto del total de secuencia fue de 0,02 %. El porcentaje de secuencia repetida encontrada usando TRF fue de 2,9 %.

3.4. Predicción y filtrado de SNPs

Los resultados obtenidos de AnalyzeSNPs fueron aproximadamente 500 000 sitios polimórficos, ya que el programa reporta todas las bases que difieren del consenso cuando su calidad suma más de 40, lo que en principio genera que casi cualquier base que difiera del consenso será interpretada como un SNP. Luego de la fase 1 de detección se obtuvieron 4991 SNPs en 1352 contigs. El filtrado resulta ser bastante eficiente (aprox. 99 % de SNPs desechados) en descartar las diferencias propias del alineamiento múltiple en un ensamblaje con secuencias de calidad variable que no corresponden a sitios

polimórficos (fig. 3.4).

Luego de la segunda fase para determinar la presencia de SNPs, se observa que a medida que la cobertura requerida es mayor, el número de SNPs putativos disminuye bruscamente, el efecto de los otros dos parámetros es comparativamente de menor impacto. Para compensar este efecto, y establecer un set de SNPs amplio sin perder certeza en la predicción, se consideraron tres categorías de SNPs: La primera, A, (Q=40,V=20,P=4) presenta 4318 SNPs en 1352 contigs (5,35 % del total de contigs), la segunda, B, (Q=30,V=20,P=6) contiene 1757 SNPs en 499 (1,97 %) y en la tercera categoría, C, (Q=20,V=20,P=8) se observan 1308 SNPs en 310(1,23 %) contigs. La frecuencia promedio de SNPs por contig presenta un rango que va desde 2,37 a 4,23 SNPs por contig.

Los resultados de ambas fases (predicción de SNPs y búsqueda de repeticiones) fueron agregados a los archivos GenBank para su visualización y filtrado (fig. 2.1).

3.5. Búsqueda de genes y anotación de secuencias génomicas

Luego del filtrado final se encontraron 956 (3,78 %) contigs que cumplen las condiciones establecidas y en total presentan 2849 SNPs de la categoría A, 1265 SNPs en B y 968 en C. El set total de 956 contigs representa la muestra en estudio de secuencias con alta probabilidad de venir de regiones duplicadas y que presentan al menos un SNP que serviría de marcador para esa región.

De la búsqueda realizada con blastX se encontraron 452 (48,5 %) contigs que presentan algún hit con e-value $< 1 \cdot e^{-}05$. Considerando solamente el mejor hit por contig se encuentran 184 genes distintos, en cambio, si se consideran los mejores 5 hits,

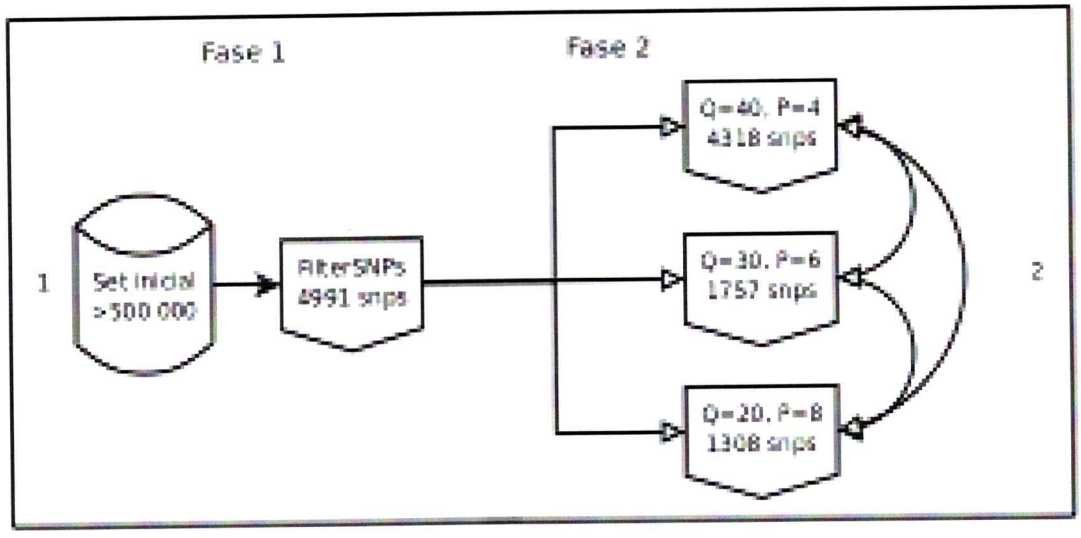

Figura 3.4: **Fase de filtrado de SNPs.** El set final de SNP se obtuvo a partir de la estrategia mostrada en la figura.

esta cifra aumenta a 469 genes distintos. Esta diferencia indica que existen múltiples anotaciones para una misma secuencia, o bien, que la proteína codificada comparte o ha compartido diversas funciones y en su secuencia génica permanece algún grado de similitud detectado por blast. Esto último se puede explicar por la presencia de genes duplicados que se han preservado de tal manera que sus funciones actuales están separadas (Ej. expresión en tejidos distintos), relacionándose por haber sido funciones alternativas de algún gen ancestral (modelo de subfuncionalización) o, que luego de un evento duplicativo hayan adquirido nuevas funciones (modelo neofuncionalización) y, por lo tanto, se debiesen considerar como secuencias parálogas.

También se observan múltiples contigs que presentan el mismo hit. *A priori* esto se explica porque los clones BAC representan una librería redundante (cobertura 18×) en la que un gen puede estar presente en distintos clones o, por la existencia de familias con

múltiples copias de algún gen (duplicación) que no fueron agrupadas en el mismo contig. Repeatmasker no identificó la totalidad de transposones presentes en las secuencias. El resultado blast indica la presencia de 66 contigs con hit a elementos transponibles, tales como transposasa o retrotransposasa, confirmando que estos elementos se encuentran ampliamente distribuidos a lo largo de las secuencias genómicas del salmón y su presencia es una fuente de polimorfismos dada por la redundancia de la secuencia de éstos. Así mismo, existen 41 contigs que presentan hit contra proteínas tipo polimerasa y 22 que presentan hit al elemento *ReO*, el cual contiene repeticiones a lo largo de su secuencia. Los SNPs presentes en este tipo de secuencias se consideran como SNP falsos repetitivos, es decir, provienen del alineamiento de transposasas localizadas en regiones distintas y que presentan diferencias nucleotídicas entre sí. Sin embargo, no es posible descartar que estos elementos se hayan originado por el mismo mecanismo de duplicación y que los SNPs encontrados en estas regiones sean del mismo tipo que el de regiones parálogas (PSVs).

En la búsqueda de genes *ab initio* realizada por FGENESH se encontraron 797 CDS, 149 sitios poly-A y 140 sitios de inicio de la transcripción. Estos resultados fueron integrados a la información obtenida en la etapa anterior y fueron visualizados para corroborar manualmente la presencia de los exones identificados previamente.

3.6. Selección de contigs para validación

Los contigs seleccionados se describen en el cuadro 3.1 y la información referente al nombre completo del gen, su largo genómico, su largo proteico y la presencia de parálogos para cada gen en humanos se resumen en el cuadro anexo A.2.

El primer contig presenta hit blast al gen *Prune* de *D. rerio* (Unigene ID number 563343). La función de esta proteína en humanos está relacionada con actividad

							SNPs		
ID	Contig	Identifier	Gene symbol	E-value	Exon hit	Identity(%)	Set [a]	Position	# predicted SNPs [b]
1	1044	563343[c]	*Prune*	$3e^{-15}$	7	36/53(67%)	A	Exon	1
2	1139	Q5RG11	*Rfx5*	$2e^{-30}$	6	66/67(98%)	C	Intron	2*
3	2664	A4QP33	*Ppp2r5c*	0.02	10	19/19(100%)	C	Intron	1
4	3478	Q5NDG1	*St6galnac4*	$5e^{-19}$	2,3,4	96/143(67%)	A	Exon	2*
5	4211	Q7ZVF4	*Rars*	$1e^{-5}$	8	24/29(82%)	A,B	Intron	1
6	4331	Q6PH56	*Hsp70*	$1e^{-63}$	1	125/173(72%)	A	Exon	2*
7	5133	B8JIF6	*Vn2r1p*	$3e^{-26}$	1,2	89/148(60%)	A	Intron	1*
8	5345	A1A5X1	*Mapk3*	$3e^{-23}$	1	57/102(56%)	A,B	Intron	1*
9	5710	Q5M7Y0	*Col4a3bp*	$1e^{-15}$	13,14	48/61(79%)	A	Intron	1
10	6459	A8DZD1	*Galnt6*	$9e^{-24}$	2	52/57(91%)	A	Intron	1*

Cuadro 3.1: **Contigs seleccionados**. Se observa el identificador único en UniProt y el identificador correspondiente al número del contig en el ensamble original. Además se observa el e-value obtenido, el exón correspondiente y la información del SNP putativo.

[a]Categoría(s) de los SNPs encontrados en cada contig.
[b]Para evitar confusión, los SNPs validados presentan un asterisco.
[c]Unigene ID de NCBI. Los demás identificadores corresponden a Uniprot

fosfodiesterasa hacia cAMP y cGMP como sustrato. Participa en la proliferación celular siendo capaz de inducir motilidad celular y actuar como regulador negativo de NME1 (Garzia y col., 2008). En humanos está localizado en el cromosoma 1 y tiene una variante paráloga en el cromosoma 9. En *D. rerio* este gen posee 8 exones y se localiza en el cromosoma 16 en una región de 26 kb. El largo de la proteína es de 447 aa como indica el cuadro A.2, el SNP putativo se ubica en la región codificante del exon 7 de su ortólogo en Danio.

La proteína codificada por la secuencia del segundo contig seleccionado se denomina RFX5 (Regulatory factor X 5) (Uniprot ID Q5RG11). Es parte de una familia de proteínas de unión a DNA que reconocen un motivo denominado caja X, el cual es parte de la región regulatoria de la expresión del complejo MHC II. Su gen se localiza en el cromosoma 1 (humanos) y presenta una variante paráloga en el cromosoma 15. Además, la secuencia codificante de RFX5 encontrada corresponde al exón número seis de su ortólogo en *D. rerio* con un porcentaje de identidad del 98 %. El único cambio en este marco de lectura es de leucina(*D. rerio*) por metionina en Salmón.

El contig tres no presentó un hit blast con e-value menor al límite establecido de corte, sin embargo, el hit encontrado corresponde a uno de los exones de la subunidad B de la proteína fosfatasa 2A (ver cuadro 3.1) de *D. rerio*(Uniprot ID A4QP33). Este exón presenta 18 aminoácidos conservados entre ambas especies por lo que en este caso el e-value no fue usado como criterio de selección. La subunidad B es parte de una familia de proteínas regulatorias que se encargan de modular la selectividad por el sustrato y la actividad catalítica de la fosfatasa 2A. La proteína fosfatasa 2A presenta dos subunidades más que se asocian con una variedad de subunidades regulatorias, entre ellas, la subunidad B. El hit representa el exón 10 de esta proteína en *Danio rerio*. El gen presenta 16 exones en una región genómica de 61kb lo que implica una gran cantidad de secuencia intrónica

ya que el largo aminoacídico es de sólo 578 aa. En humanos presenta tres variantes parálogas y nueve variantes de splicing alternativo.

El hit encontrado en el contig cuatro corresponde a una proteína de membrana del aparato de Golgi encargada de transferir un grupo de ácido siálico de CMP-N-acetilneuroaminato a distintas glicoproteínas y glicolípidos, con preferencia por estos últimos (Uniprot ID Q5NDG1). Pertenece a la familia 29 de las glicosiltransferasas. El cDNA de esta proteína fue caracterizado para *Salmo salar* presentando un largo de 295 aminoácidos (Uniprot ID B5X2G4). Sin embargo, al comparar la secuencia génomica obtenida con la del cDNA se aprecian diferencias significativas entre ellas, esto sugiere la presencia de una copia paráloga expresándose en el tejido estudiado. En *Danio rerio* esta proteína presenta 4 exones localizados en el cromosoma 21, y el contig estudiado consiste en el exón 3, 4 y parte del segundo. Como indica el cuadro 3.1 el SNP se ubica en posición intrónica entre los exones 3 y 4.

El resultado blast para el quinto contig corresponde a la proteína arginyl-tRNA sintetasa (Uniprot ID Q7ZVF4). Su secuencia está previamente descrita para salmón (Uniprot ID B5X4E3). La proteína cataliza el acoplamiento de arginina a su tRNA actuando como ligasa. El gen presenta 15 exones en una región génomica de 50kb. En danio presenta un largo de 661 aa similar al del salmón de 660 aa. En el contig la secuencia hit pertenece a parte del exón 8 del gen en *D. rerio*. En humanos este gen presenta un parálogo y 7 variantes de splicing.

En la región 6 está presente el hit blast a la proteína HSP70. La secuencia de esta proteína está disponible en múltiples organismos incluyendo trucha (Q5KT35), salmón del Atlántico (B5X4Z3) y pez zebra (Q6PH56), pertenece a la familia de las *Heat shock proteins* de 70kDa encargadas de responder a situaciones de estrés al interior de la célula. Su largo en *D. rerio* es de 643 aa, similar al observado en *O. mykiss* (644 aa)

y *S. salar* (644 aa). Presenta sólo un exón y recientemente se describió una secuencia paráloga en trucha (Ojima y col., 2005). Se encuentra altamente conservada a través de la evolución y es la principal familia de proteínas en vertebrados encargadas de actuar como chaperonas moleculares, asegurando el correcto plegamiento de las proteínas intracelulares. El locus de este gen se ubica en el cromosoma 3 del genoma de *D. rerio*. Ambos SNPs putativos se ubican en la región codificante del gen.

El gen observado en el contig 7 corresponde a un receptor transmembrana asociado a proteína G, su función está relacionada con la captación de olores y/o feromonas (Uniprot ID B8JIF6). Pertenece a la familia de receptores glutamato metabotrópicos. En *D. rerio* presenta 6 exones localizados en el cromosoma 17 que corresponden a una región genómica de 6 kb. El contig descrito alberga el primer y segundo exón y el SNP se ubica entre éstos. Su largo aminoacídico es de 837 aa con 7 dominios transmembrana. En humanos esta secuencia representa a un pseudogen ya que no se han encontrado transcritos expresados de este locus. Es conocida la poca sensibilidad humana para la captación de olores en comparación con otras especies sugiriendo que esta forma del gen (y sus múltiples copias) se ha mantenido activa en aquellas y no en el humano (transformándose en pseudogenes).

El primer exón del gen de la quinasa 3 activada por mitógeno (*Mapk3*) fue encontrado en el octavo contig (Uniprot ID A1A5X1). Esta proteína es parte de la famila de las MAP quinasas que actúan en la cascada de señalización que regula procesos celulares tales como proliferación, diferenciación y ciclo celular en respuesta a distintos estímulos provenientes del medio externo. Su gen presenta 10 exones en *D. rerio* que cubren una región de 46 kb en el cromosoma 11. El largo de la proteína es de 408 aminoácidos y para el salmón es de 406 aminoácidos. Esta proteína se encarga de fosforilar y reprimir la actividad de un factor de transcripción involucrado en la regulación de la expresión tejido

específica y diferenciación celular. En humanos presenta dos genes parálogos, *MAPK5 y MAPK2.* El SNP se encuentra 60 pares de bases río arriba del codón de inicio.

El noveno contig alberga los exones 13 y 14 del gen de la proteína de unión a colágeno tipo IV (Uniprot ID Q5M7Y0). Se encarga de mediar el tráfico intracelular de ceramida sin utilizar vesículas, se encuentra en el citoplasma, retículo endoplásmico y principalmente en el aparato de Golgi. Presenta un total de 17 exones en una región de 55 kb del cromosoma 15 de *D. rerio.* El largo proteico es de 620 aminoácidos. En humanos esta proteína está relacionada con el Sindrome de Goodpasture, enfermedad autoinmune que afecta a los pulmones y riñones. Presenta dos variantes extras por splicing alternativo y tres genes parálogos. El SNP putativo se ubica en la región intrónica rio abajo del exón 14.

En el último contig se encuentra el hit blast al polipéptido N-acetylgalactosaminyl transferasa (Uniprot ID A8DZD1). Este gen presenta 10 exones en *D. rerio*, de los cuales el segundo esta presente en este contig. Esta localizado en el cromosoma 23, su largo es de 21 kb y codifica para una proteína de largo 629 aa. Cataliza la transferencia de N-acetyl galactosamina a un residuo treonina o serina de la proteína receptora. Esta anclado a la membrana del Golgi y participa en la síntesis de fibronectina oncofetal. En humanos presenta 13 variantes de splicing y dos genes parálogos. El SNP se ubica en la región intrónica correspondiente al intrón 3 de su ortólogo en *D. rerio.*

3.7. Extracción de DNA genómico, amplificación y purificación de los contigs seleccionados

La concentración obtenida luego de la extracción del DNA genómico de salmón fue de 65 ng/uL que se diluyó a la mitad para realizar la amplificación por PCR. Los

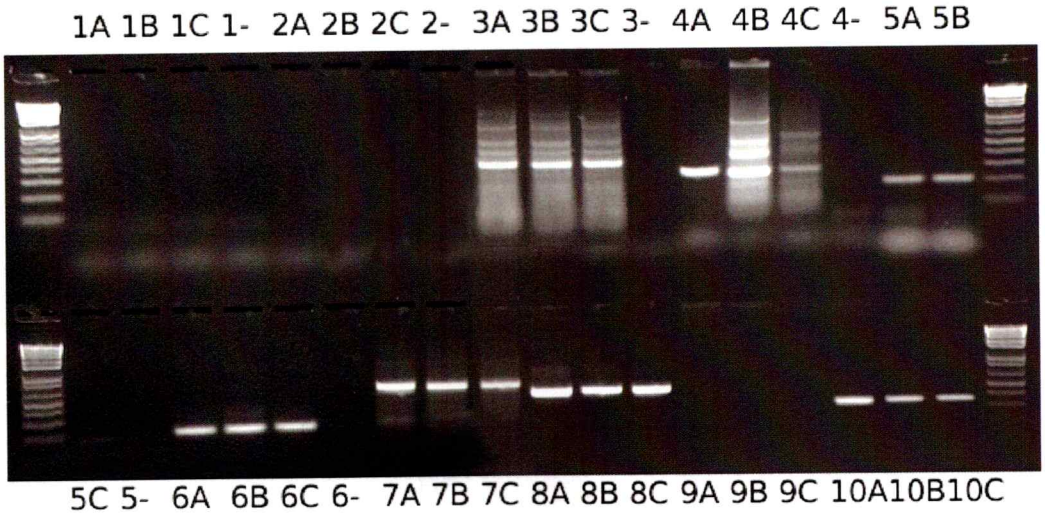

Figura 3.5: **PCR de las 10 regiones candidatas.** Electroforesis en gel para visualizar las bandas correspondientes a la amplificación por PCR de las secuencias candidatas.

partidores diseñados (cuadro A.1) fueron probados bajo un protocolo estándar de PCR para todas las muestras resultando en la amplificación correcta de 7 de las 10 regiones (fig. 3.5). Las restantes tres regiones fueron sometidas a optimización de las condiciones PCR, obteniéndose amplicones en dos de ellas. Los partidores de la región uno nunca amplificaron probablemente debido a la formación de dímeros y favorecida por el hecho que el partidor directo fue diseñado con el polimorfismo ubicado en la última base del extremo 3'. Esto disminuye la fuerza de unión entre el templado y el partidor, generando ausencia del producto. Se redujo la Ta para disminuir la especificidad de la unión sin mejores resultados, por lo que esta región se descartó de los análisis posteriores.

En las electroforesis en agarosa todas las bandas obtenidas corresponden a los tamaños esperados para la región amplificada (fig. 3.5). La purificación del producto PCR resultó en 26 muestras provenientes de cada locus amplificado para cada uno de los individuos usados (no se pudo amplificar la región 3 para el individuo A, sí para los dos

restantes.)

3.8. Secuenciación y genotipado

Los resultados de la secuenciación variaron dependiendo de la región analizada, para la secuenciación con el partidor directo 10 secuencias no pudieron ser analizadas por presentar problemas de señal/ruido en la región del SNP. Estas regiones fueron amplificadas y purificadas por segunda vez y se enviaron a secuenciar con el partidor reverso, al igual que el resto de los productos PCR. La secuenciación con el partidor reverso arrojó 10 secuencias de baja calidad señal/ruido, repitiéndose el resultado negativo para los contigs 2 y 8 (ver cuadro 3.2). Se obtuvieron 7 secuencias consenso que fueron revisadas manualmente en consed y resultaron en un 100 % de identidad con las secuencias consenso obtenidas en la etapa *in silico*, exceptuando la región 6 con un 99 %. El largo obtenido de éstas fue igual al esperado excepto para la región 7 que resultó 84 pb más largo y para la región 10 que fue de 41 pb menor. Los resultados de la secuenciación y genotipado se resumen en el cuadro 3.2.

Todos los genes encontrados presentan al menos un parálogo en humanos. Se puede suponer que la búsqueda es redundante por el hecho que si el gen ya está duplicado en humanos lo estaría también en salmón. Sin embargo, el proceso de rediploidización produce la pérdida de genes que no necesariamente son igualmente útiles para todas las especies. Los mecanismos de preservación de duplicados actúan en función de la utilidad conferida por la mutación fijada en una copia paráloga, que le permitió escapar a la selección purificadora de forma distinta en cada especie. Esto es reafirmado por la hipótesis de RGL (Reciprocal Gene loss) y está relacionado con los procesos de especiación. Visto desde otro punto de vista, la metodología utilizada es independiente de si existe un gen duplicado o no. Si se considera que estos genes se han preservado como

duplicados en todas las especies es porque todas sus copias aportan una funcionalidad necesaria para los sistemas biológicos y la estrategia simplemente identifica estos genes a partir de estas ventajas selectivas dadas por las mutaciones o SNPs.

3.8.1. Región 2 y 8

Las secuencias obtenidas de estas regiones resultaron con baja calidad señal/ruido para todos los individuos lo que impidió generar un consenso. Sin embargo, fue posible un análisis manual debido a que las secuencias individuales obtenidas fueron de largo similar al esperado. La figura 3.6 representa el extremo de las secuencias obtenidas para la región 2, se observa que a partir de la posición 190 (peak más álto) no se observa ruido de base y la secuencia observada a partir de esta posición corresponde a la del partidor reverso. Al comparar los peaks presentes antes de la posición 190 y después se puede observar que la misma secuencia se encuentra dos veces en los últimos 40 pb (posición 190 - 20 y 190 + 20). Esto se explicaría por la presencia de una deleción y/o inserción de 20pb en uno de los templados. La deleción estaría presente en todos los individuos, y se observa el mismo patrón en ambos extremos, lo que indica que se secuenciaron dos templados parálogos.

El genotipado del locus 2 utilizando el SNP putativo indica que dos de los tres individuos son heterocigotos T/C. A pesar del ruido, el hecho que uno de los individuos fuese homocigoto para este marcador permitió la identificación del SNP en esta región (fig. 3.7). Si todos los individuos hubiesen sido heterocigotos no hubiese sido posible diferenciar las variantes alélicas del SNP. La presencia de copias parálogas en esta región indica que el SNP encontrado corresponde a un MSV.

La región 8 presenta ruido en toda la secuencia proveniente del partidor directo, observándose múltiples secuencias sobrelapadas. Este problema se encuentra documen-

	Individual genotype and length			Length (bp)[a]		SNP		
Locus ID	A	B	C	Predicted	Obtained	SNP Pos.	Flanking 5'-3'	Obs.
2	208 (T/C)	210 (T/C)	208 (T/T)	212	-	123	TCTTG (T/C) AACCA	MSV
3*	-	373 (T/T)	421 (T/T)	418	418	355	TTTTT (T/T) CCCCC	-
4	202 (G/T)	200 (G/G)	196 (G/T)	287	287	56	AATTG (G/T) CATAT	Heterocigoto/MSV
5*	200 (C/C)	121 (C/C)	142 (C/C)	200	200	140	AAAAA (C/C) TCCAT	-
6	227 (T/C)	213 (T/C)	227 (T/C)	227	227	155	AAGAT (T/C) GCAGC	PSV
7	621 (T/C)	493 (T/T)	580 (T/T)	537	621	230	ATTAG (T/C) TTCGC	MSV
8	478 (G/C)	479 (?)	469 (G/C)	492	-	388	AGTCT (G/C) TAAAA	?
9*	434 (T/T)	246 (T/T)	450 (T/T)	482	482	186	ATATT (T/T) CTTAT	-
10	312 (A/G)	312 (A/A)	280 (A/A)	353	312	64	TTCAA (A/G) CGCTG	Heterocigoto/MSV

Cuadro 3.2: **Resultados de secuenciación y genotipado.** El cuadro muestra los resultados obtenidos de la secuenciación de cada región para cada individuo. Muestra además el largo obtenido, el largo esperado y la conclusión sobre el origen del SNP encontrado, en los casos con asterisco el SNP no pudo ser validado.

[a]Largo esperado y largo obtenido luego de secuenciar y obtener un consenso para los tres individuos.

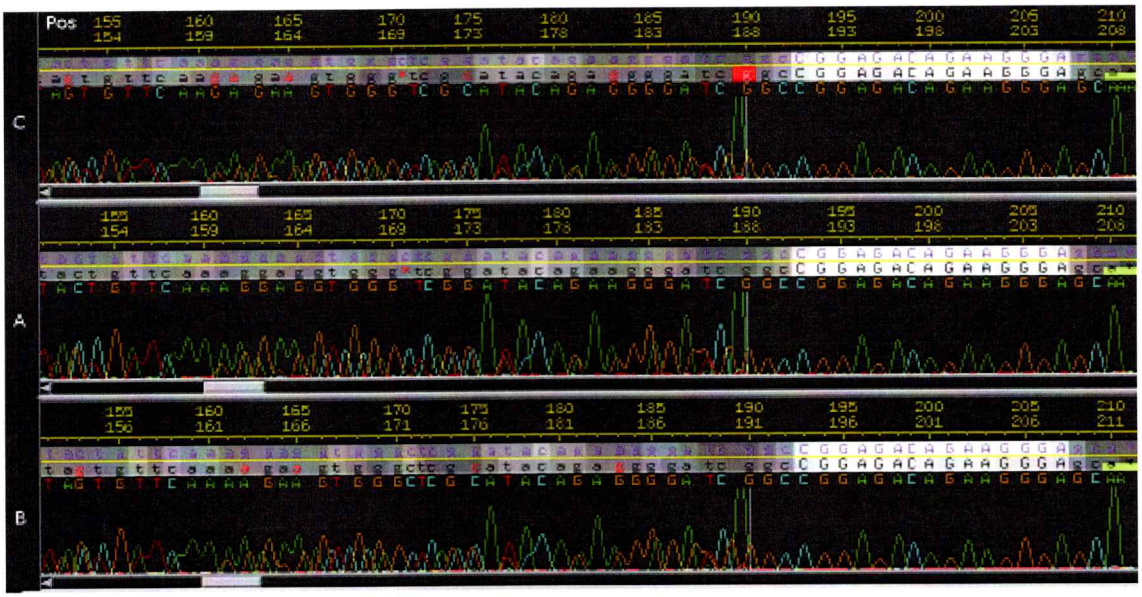

Figura 3.6: **Cromatograma en la región final de la región 2.** El área más clara muestra los 40 pb finales en cada individuo, se aprecia como se repite la secuencia del partidor dos veces en esta área y el sobrelapamiento con la secuencia proveniente de una segunda copia de templado (copia paráloga).

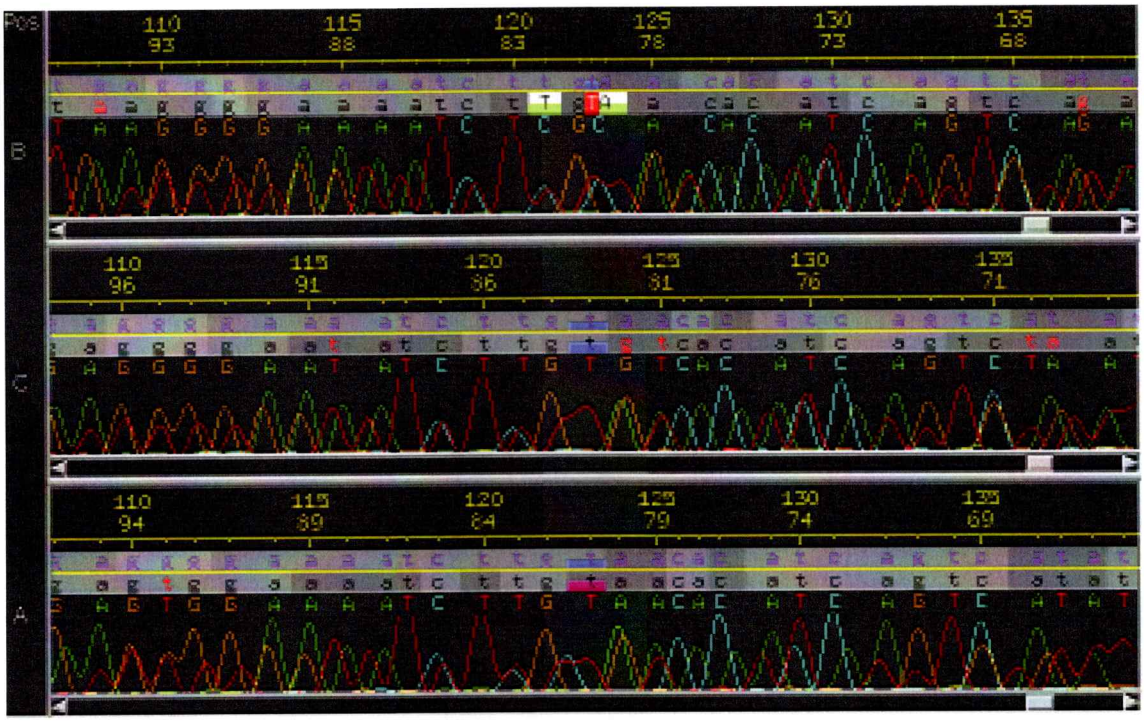

Figura 3.7: **Cromatograma en la región del SNP del contig 2.** El área gris muestra la posición del SNP predicho, el individuo A y B son heterocigotos (peaks celeste de citosina sobrelapado con peak rojo de timina). El individuo C es homocigoto T/T.

tado y se puede deber a diferentes causas: A) Las muestras estaban contaminadas con algún reactivo inhibidor. B) Exceso de partidor y C) presencia de distintas secuencias provenientes de la amplificación por PCR de regiones duplicadas de tamaño similar pero con alguna inserción o deleción presente.

La purificación del producto PCR en gel de agarosa se realizó en duplicado para ambas regiones con el fin de disminuir la probabilidad de un exceso de partidores y/o de contaminantes en las muestras. La secuencia obtenida con el partidor reverso para la región 8 fue de buena calidad hasta la presencia de una región poly-A (fig. B.3). El genotipado de los individuos A y C fue posible observando la secuencia en la posición putativa del SNP y comparando los peaks presentes. La secuencia del individuo B no permitió ningún nivel de resolución de su genotipo para el SNP interrogado producto del excesivo ruido. El ruido observado es producto del deslizamiento de la polimerasa provocado por la region poly-A (discutido para la región 7) o corresponde al sobrelapamiento de una región duplicada. Esta última explicación concuerda con la hipótesis propuesta.

3.8.2. Región 3, 5 y 9

La región 3, 5 y 9, correspondientes a la subunidad B de la fosfatasa 2A, a una Arg-tRNA sintetasa y a una proteína de unión a colágeno respectivamente, no presentaron heterocigozidad para el SNP, por lo que su validación falló. El largo de las secuencias obtenidas fue similar al esperado y su calidad era adecuada para evaluar los alelos presentes en cada individuo. Estos SNPs putativos pertenecen a tres categorías distintas diseñadas en la fase 2 del filtrado (fig.3.4, cuadro 3.1). Este filtro no mejoró la tasa de validación de SNPs sino que disminuyó el número de SNPs candidatos y permitió añadir un nivel de información extra.

Además, la secuencia que flanquea el SNP putativo (contig 3 y 5, cuadro 3.2) corresponde a una pequeña región poly-N donde el SNP se ubica al final de ésta. Estas regiones son propensas a producir artefactos de buena calidad en el ensamblaje por lo que puede ser posible filtrarlas en la búsqueda *in silico*. En una nueva versión de la metodología se incorporará un nuevo filtro que evite o que catalogue estos SNPs putativos con una menor probabilidad de ser verdaderos lo que mejorará la tasa de validación.

El caso de la región 9 resulta ejemplificador para otro tipo de situaciones dadas por artefactos del ensamblaje. El SNP putativo correspondía a una deleción de una base nucleotídica. Debido a las propiedades químicas de las tinciones usadas en las máquinas de secuenciación actuales, la probabilidad que dos bases se encuentren muy juntas o muy separadas es alta. Esto generaría secuencias de buena calidad pero que al momento de buscar SNPs, generaría falsas deleciones (o inserciones). Por lo tanto, el SNP putativo sería un artefacto del ensamblaje y no un SNP real. Para mejorar la tasa de validación de SNPs encontrados es posible incluir esta información previamente.

Sin embargo, es necesario considerar que el número de individuos usados para el genotipado fue de tres para la region 5 y 9, y de sólo dos para el contig 3. Esto indica que no podemos descartar que estos SNPs no estén presentes en la población ya que se requeriría un tamaño muestral mayor.

3.8.3. Región 4, 7 y 10

La secuencia consenso generada del contig cuatro presentó un largo similar al esperado y con 100% de identidad con la secuencia obtenida previamente. La figura 3.8 muestra la presencia de ambos alelos en la posición del SNP para los individuos A y C. El individuo B resultó homocigoto G/G. La presencia del SNP en dos de tres individuos,

indica la posibilidad de un alto grado de polimorfismo de esta variante en la población, lo que podría ser producto de regiones parálogas. Sin embargo, con la técnica utilizada no es posible determinar con certeza si este polimorfismo corresponde a un heterocigoto producto del estado diploide del genoma o a un MSV.

La región 10 presentó dificultades al momento de ser secuenciada con el partidor directo principalmente porque contenía una región poly-A de 15 pb cercana a éste. La secuenciación con el partidor reverso generó secuencias de buena calidad que conformaron un consenso 40 pb más pequeño de lo esperado. El resultado del genotipado indica que solo el individuo A es heterocigoto para este marcador (fig. B.4). Al igual que en el contig cuatro, no se puede determinar si este SNP corresponde a un MSV o no.

La secuenciación del contig 7 permitió observar la presencia de ambos alelos sólo en el individuo A (fig. B.2). También se encontró un segundo SNP en este individuo posicionado al inicio del partidor directo en la posición 23. Los individuos B y C presentaron una baja calidad en el valor de señal/ruido observado en esa posición que no permitió el genotipado.

La secuencia consenso generada para este locus presentó un largo 84 pb mayor al tamaño esperado de 537 pb (cuadro 3.2). Esto pudo haber ocurrido por la presencia de un región poly-T al medio de la secuencia que generó un salto de la polimerasa al momento de secuenciar. Esto se produce porque la polimerasa se disocia de la región poly-N y se reubica en cierta cantidad de pares de bases más adelante o atrás generando ruido y secuencias de largo variable. La otra razón es que existan múltiples copias del templado y que éstas difieran en tamaño, es decir, que hayan ocurrido inserciones y/o deleciones en alguna de sus copias. Al momento de secuenciar estas copias se sobrelapan y se observa el ruido presente en el cromatograma de la figura 3.9. Lo interesante de esta región es

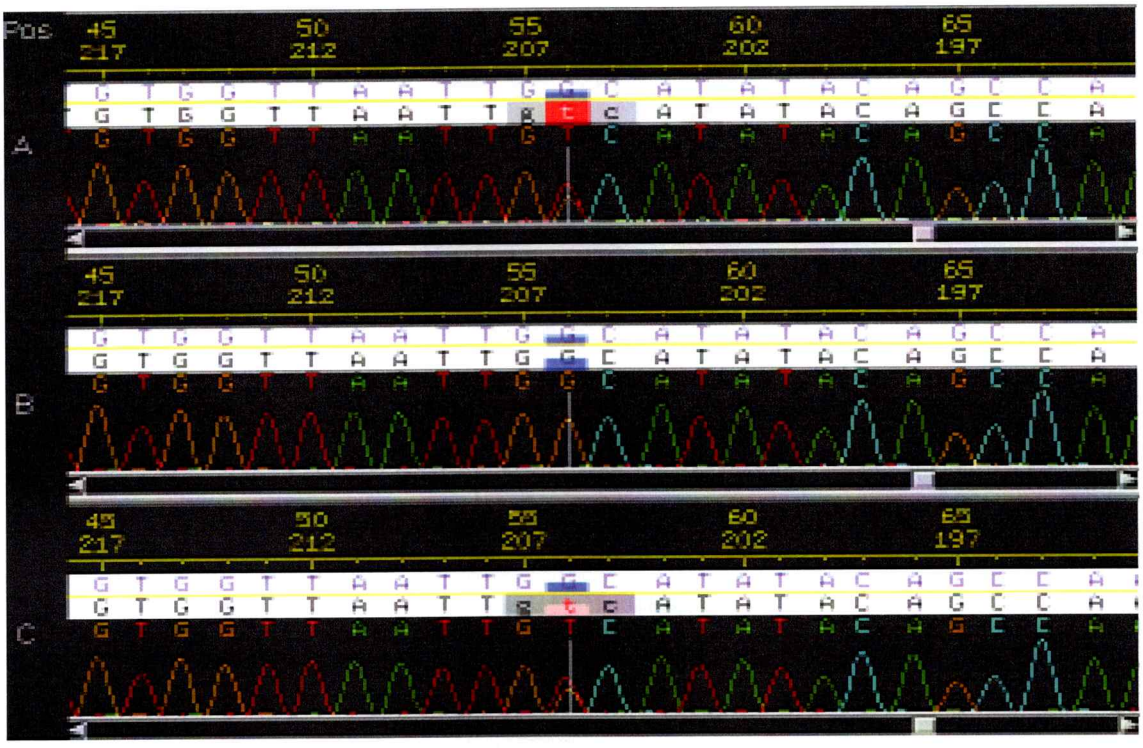

Figura 3.8: **Cromatograma en la región del SNP en el contig 4.** La posición del SNP está indicada por la linea blanca, el individuo A y C son heterocigotos (primera y tercera secuencia). El individuo B es homocigoto G/G.

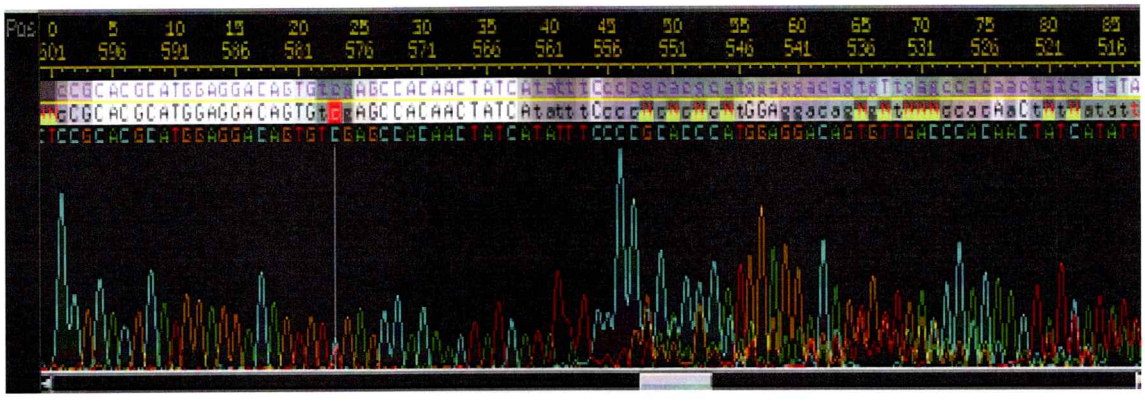

Figura 3.9: **Cromatograma en la región del extremo del contig 7.** Se observa una serie de peaks de Citosina (peaks achurados), repetidos a partir de la posición 45, para mostrar la evidencia de una inserción en una de las copias parálogas. Se observa también un SNP heterocigoto C/T en la posición 23.

que en ambos extremos existen 42 pares de bases extras al momento de alinear las secuencias en ambos sentidos (similar a lo ocurrido con el contig 2), y estos 42 pares de bases extras corresponden a la secuencia del partidor más 20 pares de bases. En la figura 3.9 se observa como se repite la secuencia del partidor en el extremo a partir de la posición 46 (los peaks más altos representan el inicio de la secuenciación) resultando en una secuencia más larga que otra. Esto sugiere que la secuencia se encuentra duplicada y que una de sus copias sufrió una inserción o deleción. Esta evidencia indica que el SNP presente corresponde a un MSV, y está asociado al intrón 2 del gen del receptor olfatorio VN2R1P.

La duplicación génica de receptores olfatorios y su organizacion génomica han sido estudiados. Se estima que en humanos y ratones existen más de 1000 genes que codifican para estas proteínas y en *D. rerio* existen alrededor de 100 (Kratz y col., 2002). Es interesante notar que, en contraste con el ratón y el pez zebra, en los humanos una gran proporción son pseudogenes. La pérdida de estos duplicados en humanos se atribuye a una dispensabilidad de algún tipo particular de quimiorecepción. Además, se ha propuesto que los salmones utilizan este tipo de receptores para retornar, luego de sus movimientos migratorios, a sus lugares de procedencia original, y que por ende mantienen una alta proporción de estos genes como duplicados funcionales.

3.8.4. Región 6

La secuencia obtenida para esta región fue de 227 pares de bases y con buena calidad para todos los individuos. La secuencia obtenida se comparó con la secuencia cDNA previamente caracterizada (UniProt ID B5X4Z3) resultando en 7 diferencias aminoacídicas. Estas diferencias pueden estar originadas por la naturaleza del tejido desde donde se obtuvo la secuencia, indicando que existe una variación intragénomica.

Los polimorfismos pueden estar dados por la presencia de múltiples parálogos funcionales que difieren en su expresión y localización. Estos han sido optimizados para actuar en tejidos distintos o en distintas etapas del desarrollo (subfuncionalización) luego de haber pasado por un intenso período de selección.

En la secuencia investigada se encontraron nueve SNPs. Siete de aquellos resultaron heterocigotos para todos los individuos (fig. B.1). Sólo uno de estos 7 SNPs corresponde a una mutación no sinónima producto de una transversión de guanina por citosina, que se traduce en el cambio del aminoácido 554 de serina por treonina en la proteína HSP70. El cambio ocurrido no debiese afectar su función debido a que las propiedades químicas de ambos aminoácidos son similares.

Sorprendentemente, sólo dos de estos SNPs estaban presentes en el set inicial de SNPs putativos predichos. Resulta improbable que en el corto tiempo evolutivo de divergencia se hayan acumulado tantas diferencias nucleotídicas entre los mismos genes de individuos distintos de la misma especie. Esto sugiere que inicialmente se encontraron SNPs producto de la diferencia en la secuencia de dos copias parálogas (dos clones distintos) y que realmente existen más de dos. Una posible explicación es que la enzima de restricción no cortó en el lugar donde se encontraban los otros parálogos ya que presentaban una secuencia distinta en el sitio de corte, al ocurrir esto, no sería posible alinearlos juntos y, por lo tanto, predecir estos polimorfismos *a priori*. Para realizar el alineamiento posterior se consideran sólo como dos parálogos (ssal a y ssal b) al igual que los descritos previamente en *O. mykiss*.

La presencia de sólo un individuo heterocigoto para dos de los SNPs, indica además que la presencia de sitios polimórficos entre individuos de la especie sigue estando presente en regiones duplicadas, confirmando la presencia de MSVs. Además, el resultado está en concordancia con lo propuesto anteriormente por Hayes y col. (2006),

Figura 3.10: **Vista en consed de las lecturas obtenidas para la región 6.** Se observa la posición relativa al consenso de 6 de los polimorfismos encontrados. Los SNPs en la posición 80 y 110 corresponden a MSVs ya que presentan sólo dos individuos homocigotos (color azul bajo la base). El resto de los SNPs corresponden a PSVs ya que todos los individuos son heterocigotos (color rosado bajo la base).

en que la densidad de SNPs en regiones parálogas (1 SNP cada 28 pares de bases) se desvía ampliamente del promedio calculado para *Salmo salar* (1 SNP cada 600 pb) y humanos (1 cada 1000 pb).

3.9. Alineamiento *Hsp70* y consideraciones finales

La región con mayor evidencia de presentar múltiples copias parálogas es la correspondiente al gen *Hsp70* debido a la presencia de múltiples SNPs del tipo PSVs. En el caso del resto de las regiones secuenciadas, donde el SNP estaba localizado en secuencia intergénica (intrón, UTR) el número de heterocigotos obtenidos en relación con el número de individuos genotipados es variable, por lo que este criterio no es suficiente para determinar duplicación. Es interesante notar que las mutaciones que ocurran en estas regiones, en uno de los duplicados, no serán fijadas completamente en el genoma

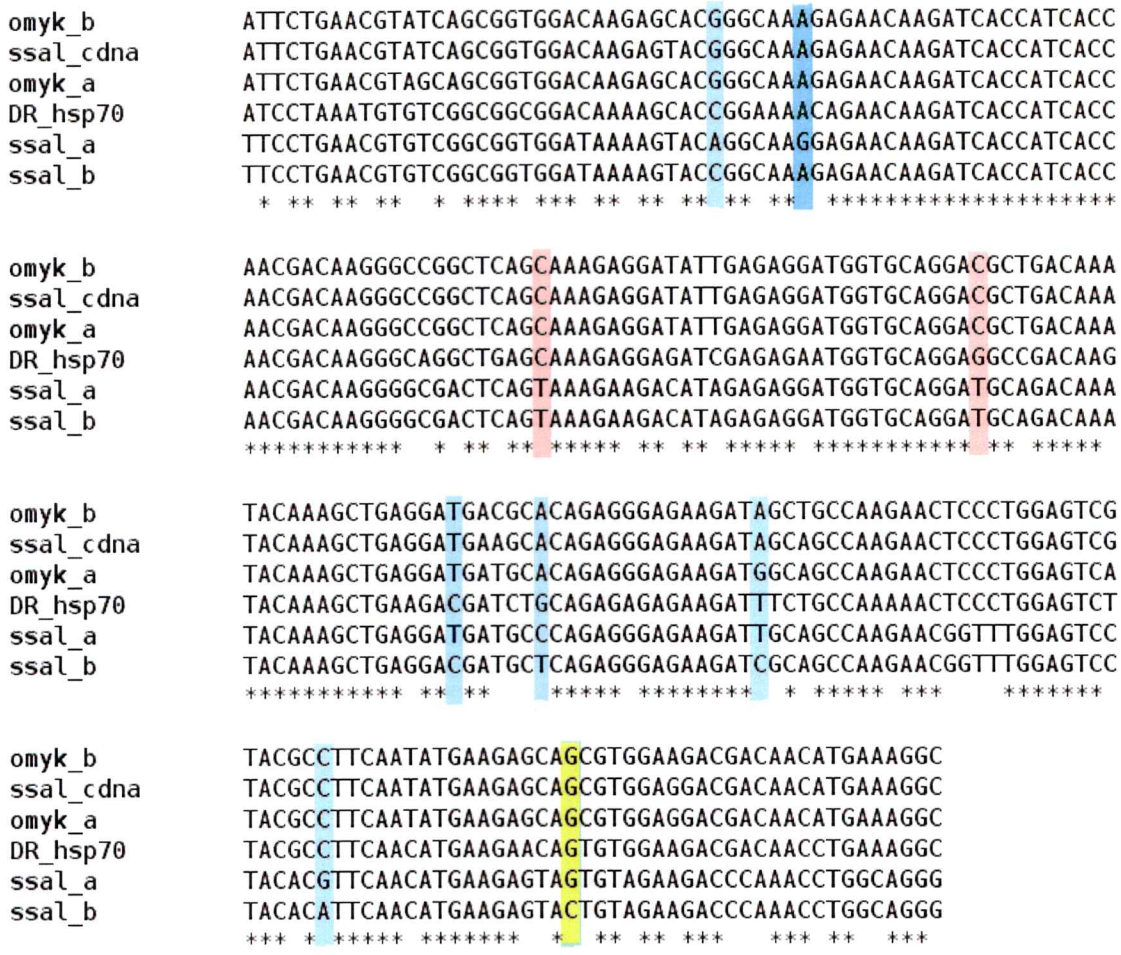

Figura 3.11: **Alineamiento de parte del gen *Hsp70* correspondiente a la región secuenciada.** El alineamiento muestra los 9 SNPs encontrados. En celeste se encuentra la posición de los PSVs sinónimos, en amarillo la posición del PSV no sinónimo y en rojo los SNPs que no resultaron heterocigotos en todos los individuos (MSVs). El alineamiento muestra también los sitios conservados entre el gen *Hsp70* de *D. rerio* (DR hsp70), del cDNA previamente caracterizado para *S. salar* (ssal cdna), de los genes parálogos en *O. mykiss* (omyk a y omyk b) y de las dos secuencias putativas de la parte de gen secuenciado para salmón en este estudio.

y así, muchos individuos presentarán sólo un alelo debido a que la fijación de éstos está relacionada con la funcionalidad que le entrega la mutación.

También es probable que las regiones no codificantes no sólo hayan acumulado una o dos mutaciones, si no que pueden haber sufrido una divergencia mayor que no sería detectada por similitud nucleotídica en el ensamblaje. Sin embargo, la divergencia presente entre los duplicados en el genoma del salmón es menor debido a que el evento es relativamente reciente.

El alineamiento de parte del gen *Hsp70*, que se muestra en la figura 3.11, muestra las diferencias y las partes más conservadas entre tres especies. La edad de los duplicados se puede estimar partiendo de la base que las mutaciones neutrales ocurren a una misma tasa mutacional para ambos duplicados en cualquier especie salmónida. Las diferencias se pueden calcular a partir de la razón entre el Ka y Ks.

El cálculo de la tasa de sustituciones sinónimas Ks y de sustituciones no sinónimas Ka para las secuencias alineadas en la figura 3.11, y la razón entre estos dos valores (Ka/Ks) como medida de la selección (cuadro anexo A.4) indica que los parálogos de *Hsp70* de *O. mykiss* han sufrido más variaciones sinónimas en comparación con los parálogos en salmón del Atlántico respecto al ortólogo único de Danio (ver cuadro anexo A.4). En el modelo neutral esto indicaría un tiempo de divergencia mayor, lo que no es posible bajo la hipótesis aceptada que ambos pares de parálogos se originaron en el mismo evento de duplicación. Esto indica que la variación en sitios silentes no siempre se ajusta a la teoría neutral, ya que es posible que exista una preferencia por el uso de ciertos codones distinto en cada especie. También, se ha observado en la literatura que usualmente cuando el número de sustituciones aminoacídicas (Ka) varía entre parálogos, el valor de Ks lo hace en una medida similar. Esto se aprecia en ambas copias, tanto en trucha como en salmón, confirmando la apreciación que los sitios sinónimos también

están bajo selección purificadora.

No obstante, el valor de Ks para cada par de parálogos entre sí (variante 'a' respecto a 'b' en ambas especies) es de 0,14 lo que indica que ambos sitios han acumulado mutaciones sinónimas bajo la misma tasa de sustituciones, a partir del momento de divergencia de estos parálogos. Esto es consistente con la teoría neutral y permitiría estimar el tiempo de divergencia de los parálogos con la obtención de datos más precisos y un mayor número de genes duplicados.

A pesar que estas observaciones son correctas a nivel génico, algunos autores (Han y col., 2009) han propuesto que la selección positiva sólo actúa sobre un pequeño grupo de aminoácidos en un período corto luego de la duplicación. Considerando esto, es probable que la tasa Ka/Ks medida no sea representativa del gen completo ya que puede que los aminoácidos que han cambiado a lo largo de la evolución se encuentren en otros sitios de la proteína no analizados.

Para la validación de un SNP encontrado *in silico* es necesario que sólo un individuo presente la condición heterocigota. Los estudios utilizan múltiples individuos para determinar una frecuencia alélica en la población y así darle un uso práctico al marcador. El uso de un número bajo de individuos limita las posibilidades de éxito de la validación, debido a que la condición mínima para que se considere como SNP es que el alelo menos frecuente se encuentre en un 1 % de la población, es decir, es necesario al menos 100 individuos para descartar la validez de un SNP. En el presente estudio se validaron 6 de 9 SNPs predichos *in silico* (una tasa de éxito de 66 %) utilizando sólo tres individuos. Este valor resultaría sorprendente si se considera que estos SNPs corresponden a SNPs reales y no variantes parálogas, pero, como se ha descrito en la literatura, los SNPs encontrados en regiones duplicadas presentan una tasa de heterocigozidad mayor. Por lo tanto, la probabilidad de encontrar individuos heterocigotos (independiente del número

de individuos genotipados) es a su vez, mayor.

Esto último, más la evidencia encontrada en cada cromatograma analizado para cada región, indica que la estrategia utilizada resultó efectiva al momento de capturar duplicados debido principalmente al éxito con que se validaron los sitios polimórficos reportados en el ensamblaje. De esta forma, es posible generar una base de datos con SNPs que corresponden realmente a PSVs o MSVs y los cuales deben ser descartados para el uso como marcadores en estudios de mapeo o asociación. Finalmente, se debiese generar una lista exahustiva de este tipo de polimorfismos para todas las especies y, así, determinar el estado de variación intragenómica presente en cada una.

4. Conclusiones y perspectivas

4.1. Genoma del salmón

El análisis de las secuencias BES del genoma del salmón permitió determinar que el porcentaje A+T en esta especie es de 58,1 % y de G+C es 41,9 %. Las secuencias obtenidas representan un 4,9 % del genoma total del salmón, de este porcentaje, un 15 % se ensambló en contigs que fueron utilizados para el análisis posterior. El ensamble arrojó 25277 contigs con un largo promedio de 895 pb que representan un 0,75 % del genoma total del salmón.

El genoma del salmón presenta un 32 % de secuencia proveniente de elementos repetitivos al utilizar la base de datos de salmónidos. Con las bases de datos de Fugu y *D. rerio* este porcentaje diminuye a 5 % y 9 %, respectivamente. Esto sugiere una expansión importante de estos elementos asociado al evento de duplicación sufrido por las especies salmonídeas. La mayor diferencia entre estas especies se da para los DNA transposones que corresponden a casi un 16 % del total en comparación con lo obtenido usando *D. rerio* de un 4 %. El número de microsatélites, en el que se pueden diseñar partidores, fue de 1266, siendo los que presentan una secuencia de dinucleótidos repetida los de mayor presencia.

A priori no es posible documentar el estado duplicado del salmón, sin embargo, de

los contigs obtenidos existe un 0.5% que presentan una cobertura mucho mayor a lo esperado (>17), probablemente proveniente de segmentos duplicados en tandem con un alto número de copias, o a la presencia de un gran número de transposones. El ensamble permite observar diferencias nucleotídicas o SNPs, a partir de los contigs que presenten 4 lecturas o más (parámetro establecido en la metodología), que en total representan 5211 secuencias.

4.2. Selección de contigs y validación experimental

Utilizando el pipeline bioinformático propuesto fue posible generar un set de contigs con secuencias genómicas provenientes de un único individuo de salmón y predecir *a priori* un set de SNPs de forma que pudiesen ser validados y genotipados en distintos individuos. Se encontraron 4991 SNPs en 1352 contigs, de los cuales sólo 452 presentaron un hit blast significativo. Estos contigs corresponden a regiones duplicadas candidatas que albergan SNPs "falsos"denominados PSVs o MSVs.

Los partidores diseñados amplificaron 9 de las 10 regiones indicando que la metodología usada para el diseño de partidores fue altamente efectiva (90%). El resultado de la secuenciación arrojó 7 secuencias de buena calidad que pudieron ser comparadas con la secuencia consenso previamente descrita. Sin embargo, esto no impidió el genotipado de todos los SNPs permitiendo validar la presencia de aquellos en 6 de las 9 regiones secuenciadas (tasa de validación de 66%).

El resultado indica que 6 de estos contigs corresponden a regiones duplicadas. El set inicial comprendió 10 locus distintos para los cuales se diseñaron partidores para amplificar y secuenciar cada uno en tres individuos no emparentados. Todos los contigs seleccionados presentaron un hit blast con algún gen conocido y sorprendentemente todos ellos se encontraban duplicados en humanos.

La presencia de regiones poly-N y deleciones fueron la principal causa de la falla en la validación de los SNPs. En una nueva versión del pipeline se incluirán los filtros que consideren ambos parámetros.

4.3. Región 6, gen *Hsp70*

Una de las regiones, el contig 6, presentó 9 SNPs y con la mayoría de ellos en estado heterocigoto para cada uno de los individuos. Esto es ejemplo de PSVs, variantes de secuencias parálogas, donde las diferencias nucleotídicas observadas corresponden a dos genes parálogos de secuencia similar ubicados en distintas posiciones y que sólo difieren en estos nucléotidos. La acción de la presión selectiva sobre la proteína HSP70 produce que la mayoría de estas diferencias se encuentren en posiciones sinónimas. El patrón observado es característico de genes que no han sufrido mucha divergencia entre sí, ya sea por un corto tiempo evolutivo de separación o por una fuerte acción de la selección.

Es necesario identificar si las versiones encontradas del gen *Hsp70* representan diferentes funciones siendo expresadas ubicuamente, o representan la optimización de alguna subfunción del gen ancestral, expresándose en distintos tejidos o en distintas etapas del desarrollo. En el caso de este gen, ambas copias se han preservado porque alguno de los SNPs acumulados en ellas les ha conferido alguna propiedad que les permitió escapar a la selección purificadora y, por ende, se han preservado como duplicados funcionales a lo largo del tiempo. Se puede especular que uno de estos SNPs es el encontrado en la posición 554 que cambió el aminoácido serina por treonina debido a que es el único cambio no sinónimo observado. Esto es poco probable debido a que las propiedades químicas de ambos aminoácidos son similares y deben existir múltiples SNPs en posiciones que no fueron exploradas a lo largo de la secuencia del gen.

El genoma del salmón puede llegar a presentar 8 copias de cualquier gen debido a las 4 rondas ocurridas. El presente trabajo muestra que al menos existen más 2 y que esas copias se encuentran funcionales, lo que concuerda con la hipótesis mencionada.

Los altos valores de Ks y un valor de Ka/Ks calculado $<< 1$ evidencian que a pesar de que las copias del duplicado *Hsp70* han escapado a los mecanismos de selección (no se han convertido en pseudogenes), ambos loci siguen sujetos a la selección purificadora que se encarga de eliminar las mutaciones deletéreas de la población y sólo permite mutaciones sinónimas que no afectan la secuencia de la proteína.

4.4. Identificación de regiones duplicadas

Existen múltiples métodos y estrategias para la identificación de regiones parálogas basados en arboles filogenéticos y homología, pero, a la fecha, ninguno ha combinado el ensamble de secuencias genómicas con la búsqueda de SNPs en un sólo individuo. Todos los modelos actuales de preservación de genes duplicados están basados en la adquisición de mutaciones que le confieran una propiedad que les permita escapar a la selección purificadora. Estas mutaciones fijadas pueden ser a nivel génico o intergénico y pueden ser identificadas como SNPs al momento de alinear secuencias parálogas provenientes de un mismo individuo.

Las metodologías actuales en la búsqueda de SNPs en salmones han tratado de solucionar la presencia de SNPs duplicados (MSVs o PSVs) intentando generar partidores más específicos (método IPEC) o realizando validación cruzada de SNPs provenientes de diferentes librerías de ESTs. Con el primero, los resultados indican que la validación mejora sustancialmente pero la estrategia no permite diferenciar si la región que es amplificada esta duplicada o no, por lo que los SNPs encontrados pueden ser realmente MSVs o PSVs. En la segunda, ambos set de datos presentaron alrededor

de 7000 SNPs, luego de la validación cruzada este número se redujo a 800, es decir, sólo un 11 % de los SNPs putativos iniciales. La estrategia empleada en este trabajo representa una alternativa atractiva para descubrir este tipo de variantes y mejorar la tasa de validación de SNPs reales.

La presencia de presión selectiva en parálogos esta exenta de los problemas que acarrean los test comparativos en términos de estructura poblacional y procesos demógraficos (Nielsen, 2005). En este caso, la población esta representada por las múltiples copias del gen en un único individuo. Si bien este individuo esta bajo los efectos de ambos procesos, la variación observada es producto de la fijación de las mutaciones producidas al interior de su genoma y no entre genomas de distintos individuos. Además, es necesario desarrollar métodos estadísticos apropiados para la detección de fuerzas selectivas que actúan en regiones no codificantes para obtener una visión mas amplia de cómo están actuando estas fuerzas sobre el organismo.

4.5. Variación intragenómica

Uno de los objetivos de este trabajo es complementar la información dada por la presencia de los SNPs con sus consecuencias funcionales en regiones duplicadas. El salmón fue elegido por ser una de las especies más representativas de la familia salmonidae, que sufrió un evento de WGD relativamente reciente, lo que debiese reflejarse en una menor divergencia evolutiva y, por ende, una mayor precisión en identificar las mutaciones que llevaron a estos duplicados a preservarse. Además, sería interesante comparar genes candidatos en especies salmonídas que presenten resistencia a cierto tipo de enfermedades (ej. Virus ISA) y determinar si existen SNPs ligados a estos rasgos en loci únicos o parálogos. Por ejemplo, es probable que el salmón coho presente variantes, ya sea a nivel de parálogos o genes únicos, que den cuenta de

una resistencia mayor al virus ISA, y que puedan ser clasificados como PSVs, MSVs o SNPs normales a través de la metodología presentada en esta memoria.

A partir de esto, es importante incorporar el concepto de variación intragenómica individual. A nivel de DNA existe variación entre especies, subespecies, poblaciones e individuos. Es necesario plantearse otro nivel de variación donde en un mismo individuo existen dos proteínas que realizan funciones similares, y cuyos genes no provienen de la misma localización cromosómica, pero que en algún momento compartieron la función ancestral, probablemente ubicados bajo un mismo locus que sufrió una duplicación. Estos genes se han especializado para actuar en distintos tejidos o en distintas etapas del desarrollo y a nivel de secuencia no representan variabilidad entre individuos sino que simplemente representan la variación presente en su propio genoma. El gen *Hsp70* y sus copias parálogas representan un ejemplo de ésto.

4.6. Proyecciones

Los duplicados encontrados presentan una alta similitud a nivel de secuencia lo que permitiría utilizar la técnica FISH para generar sondas y observar su número y localización en distintos cromosomas. Para esto se requiere seleccionar un mayor número de SNPs para genotipar y así calcular la razón Ka/Ks entre los duplicados putativos. De esta forma se podrá identificar los parálogos que están bajo una fuerte presión selectiva como en el caso del gen *Hsp70* y que, por ende, se encuentran altamente conservados.

El uso de SNPs como marcadores genéticos esta sujeto a la necesidad de determinar si estos provienen de regiones duplicadas. La estrategia *in silico* propuesta considera una validación cruzada de los SNPs encontrados bajo el ensamble de secuencias de un sólo individuo con los SNPs encontrados en múltiples individuos. Por ejemplo, en estudios de

mapeo o ligamiento es necesario filtrar los SNPs encontrados para quedarse sólo con los SNPs reales que provengan de regiones únicas.

Entonces, una de las proyecciones más importantes de este trabajo es la posibilidad de diferenciar SNPs producto de la variación intergenómica, dada para una población determinada de individuos de cualquier especie, de la variación intragenómica proveniente de un sólo individuo de esa especie. De esta forma, la aplicación del pipeline descrito en el presente trabajo, denominado ParalogIT^R, se podrá combinar con la predicción de SNPs en múltiples individuos, y así, generar dos set de SNPs donde los polimorfismos compartidos en ambos sets se clasifican como SNPs duplicados (PSV o MSV). Sólo los SNPs reales serán utilizados para estudios de asociación o mapeo (Fig. 4.1) y los SNPs duplicados podrán ser identificados para estudios de evolución y divergencia.

Con las tecnólogias de secuenciación actuales, donde cada vez es menos costosa la secuenciación de un genoma individual, la estrategia propuesta representará un paso obligado en la predicción de SNPs *in silico*. Cada vez será mas fácil diferenciar variación intragénomica de las variantes alélicas presentes en distintos individuos de una especie y se podrá realizar un catálogo individual de los SNPs presentes. Una vez las secuencias del genoma del salmón estén a disposición pública, esta estrategia podrá ser aplicada para generar una base de datos exahustiva de loci duplicados respecto a las sustituciones presentes en ellos. Además, se podrán estudiar a fondo estas sustituciones para determinar cuáles de ellas son las responsables de la preservación de los duplicados funcionales, y así establecer las variantes relevantes a la evolución de los organismos, basáda en el modelo propuesto por Ohno (1970) 40 años atrás.

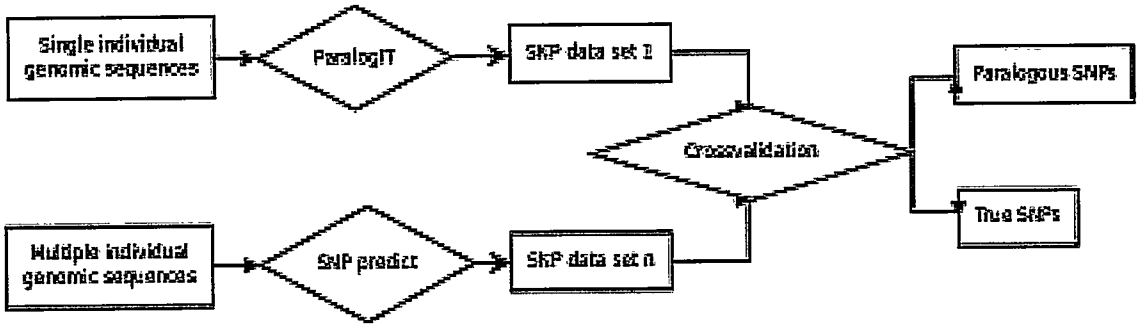

Figura 4.1: **Pipeline propuesto para la identificación de SNPs parálogos y reales.** La figura muestra paso a paso la obtención de un set de SNPs proveniente de la búsqueda en las secuencias genómicas de un único individuo y de múltiples individuos para la identificación final de SNPs reales o variantes provenientes de regiones duplicadas.

A. Cuadros

ID	Gene symbol	Length	Tm *C (F/R)	%GC(F/R)	Ta optima *C	Partidores
1	*Prune*	239	62.3/64.9	56.52/66.67	-	F 5-AGCTCGGAACCAAGCAAACGCTC-3
						R 5-ATGGGGCTCACCGCGTTCGTG-3
2	*Rfx5*	212	64.7/64.6	66.67/71.43	59	F 5-TGCCGCGATGGCAGGTGAGTC-3
						R 5-GCTCCCTTCTGTCTCCGGCCC-3
3	*Ppp2r5c*	418	63.5/60.6	63.64/54.17	58	F 5-AGTGGGTCCCTTGAGCTGGTGG-3
						R 5-ACTGGTTCAGTGAGAGTGCTGTCC-3
4	*St6galnac3*	287	63.6/59.9	63.64/48.15	60	F 5-GCAACACTGTGATCGGCACGGC-3
						R 5-TGATCGACGATAACTTCTGCAGGTGAG-3
5	*Rars*	200	64.9/60.2	70.00/48.00	58	F 5-TCTGACGCCTCACGCCTGCC-3
						R 5-ACACAGATGAAGTTGTGGCATTGGC-3
6	*Hsp70*	227	64.3/61.6	66.67/59.09	57	F 5-TCCTGAACGTGTCGGCGGTGG-3
						R 5-TCCCTGCCAGGTTTGGGTCTTC-3
7	*Vn2r1p*	575	63.2/62.8	66.67/61.90	57	F 5-CCGCACGCATGGAGGACAGTG-3
						R 5-TGATGCAGCTGCTGTGGGTGG-3
8	*Mapk3*	492	61.2/62.6	51.85/61.90	57	F 5-GGCATCAGGACTCAAACTCTGAGCAAC-3
						R 5-ATGCGTGCCCTTAGCGGTGTC-3
9	*Col4a3bp*	482	64/63.7	63.64/66.67	58	F 5-AGGTGCACCATCCCTGGACTGC-3
						R 5-GCAGGCCACACCCTCTGCAAC-3
10	*Galnt6*	353	65.2/61.8	56.52/66.67	57	F 5-ACCACAGTCCTGCACTGAAAGGC-3
						R 5-TCATGGCACACCAGCAGGCCG-3

Cuadro A.1: **Partidores.** El cuadro muestra la secuencia 5' - 3' para los partidores utilizados, la temperatura de hibridación óptima, la Tm y el largo del fragmento esperado.

Gen name	Length			Chrom.[a]	# exon[b]	# Par.[c] (HS)
	Genomic (kb)	Prot. (aa) DR	Prot. (aa) SS			
Protein prune homolog	26.23	447	-	16	8	1
Regulatory factor X 5	6.97	584	-	19	9	1
Serine/threonine-protein phosphatase 2A regulatory subunit	61.43	578	-	20	16	3
Alpha-N-acetylgalactosaminide alpha-2.6-sialyltransferase 3	7.58	291	295	21	4	3
Arginyl-tRNA synthetase (Arginine-tRNA ligase)	50.95	661	660	21	15	1
Heat shock 70 kDa protein	3.37	643	644	3	1	2
Vomeronasal type-2 receptor 1 precursor	5.97	837	-	17	6	Pseudogen
MAP kinase-activated protein kinase 2	45.39	408	406	11	10	2
Collagen type IV alpha-3-binding protein	54.67	620	-	5	17	3
Polypeptide N-acetylgalactosaminyltransferase 6	21.01	637	-	23	10	2

Cuadro A.2: **Genes presentes en los contigs seleccionados**. Se observa el nombre completo del hit encontrado en cada contig y las características de su contexto genómico en DR. Para el cálculo de Ka/Ks (cuadro A.4) se utilizó la secuencia de cDNA del gen disponible en SS. DR (*Danio rerio*), SS(*Salmo salar*), HS(*Homo sapiens*).

[a]Localización del gen en DR.

[b]Número de exones presentes en DR.

[c]Número de parálogos en HS. Fuente Ensembl o Pseudogen.org.

Repeat element	Specie DB		
	D. rerio (%)	*T. rubripes* (%)	Salmonids (%)
Retroelements	2.73	1.15	3.68
-SINEs	0.11	0.01	0.54
-LINEs	1.79	0.58	1.79
-LTRs	0.83	0.57	1.35
DNA transposons	4.3	1.98	15.78
Unclassified	0	0.01	4.61
Total interpersed	7.02	3.14	24.07
Simple repeats	1.06	1.15	1.12
Low complexity	0.75	0.77	0.65
Small RNA	0	0	0.01
Minisatellites	0	0	0.03
Total repetitive DNA	8.83	5.05	32.77

Cuadro A.3: **Elementos repetitivos.** El cuadro muestra el porcentaje presente de estos elementos a lo largo de los contigs obtenidos, se observa el porcentaje utilizando tres bases de datos distintas.

Sequence	Ka	Ks	Ka/Ks	P-Value(Fisher)
omykb - ssala	0.0477	5.4163	0.0088	$7{,}6e-40$
omykb - ssalb	0.0550	5.4496	0.0101	0
omyka - ssala	0.0591	4.9801	0.0119	$2{,}82e-38$
omyka - ssalb	0.0701	4.3719	0.0160	$5{,}12e-36$
DRhsp70 - ssala	0.0887	5.2485	0.0169	0
DRhsp70 - ssalb	0.0951	5.1201	0.0186	0
omyka - DRhsp70	0.0624	2.6548	0.0235	$9{,}11e-31$
omykb - DRhsp70	0.0471	1.7382	0.0271	$2{,}19e-25$
ssala - ssalb	0.0052	0.1430	0.0366	$5{,}13e-06$
omykb - omyka	0.0124	0.1491	0.0829	0

Cuadro A.4: **Valores y razón entre Ka y Ks.** El cuadro muestra los resultados obtenidos del cálculo de Ka y Ks para cada par de genes obtenidos de *salmo salar* (ssal a y b) y *Onchorhyncus mykiss* (omyk a y b) y su ortólogo en *Danio rerio* (DRhsp70).

B. Figuras

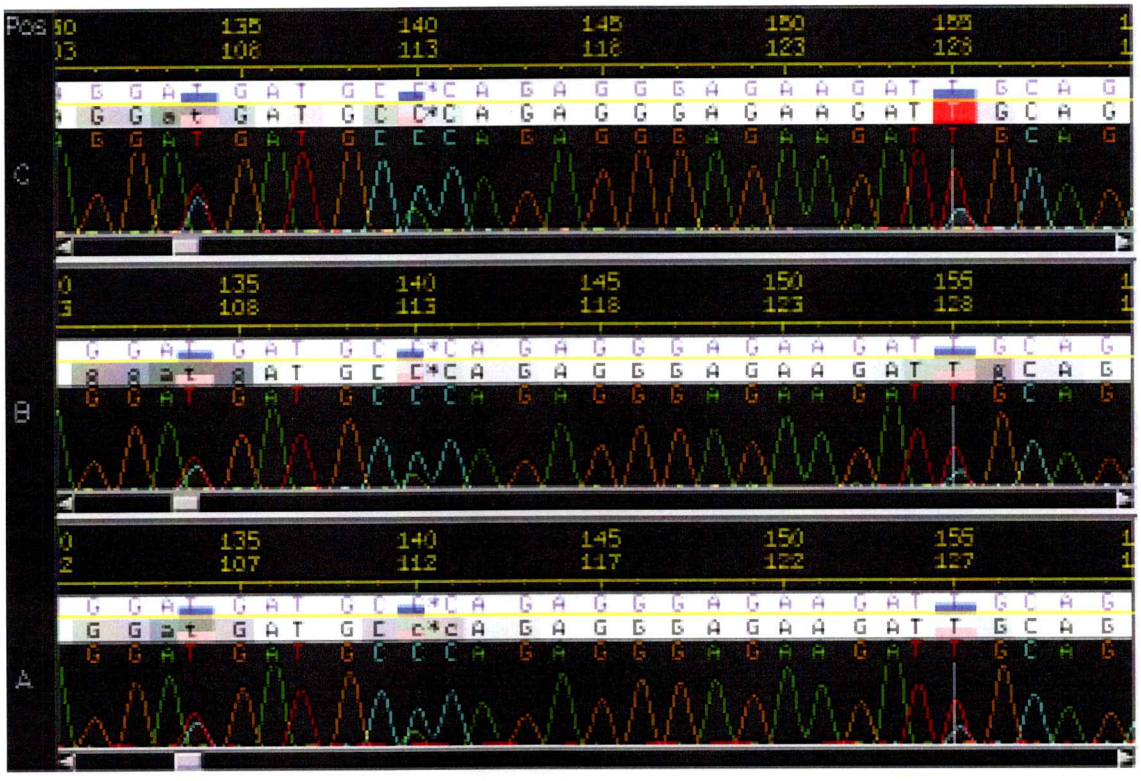

Figura B.1: **Cromatograma en región 6.** Se muestra la presencia de distintos SNPs en distintas posiciones a lo largo de la secuencia mostrada.

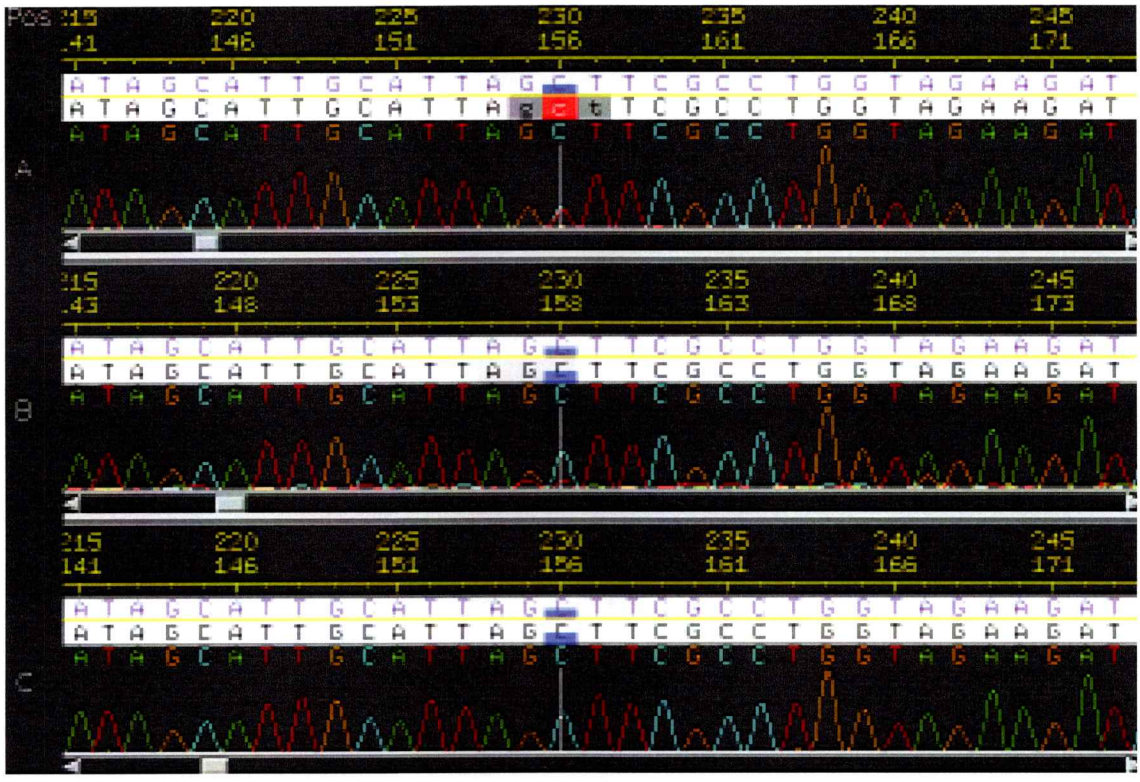

Figura B.2: **Cromatograma en región del SNP del contig 7.** La figura muestra la posición del snp y los individuos heterocigotos para este marcador.

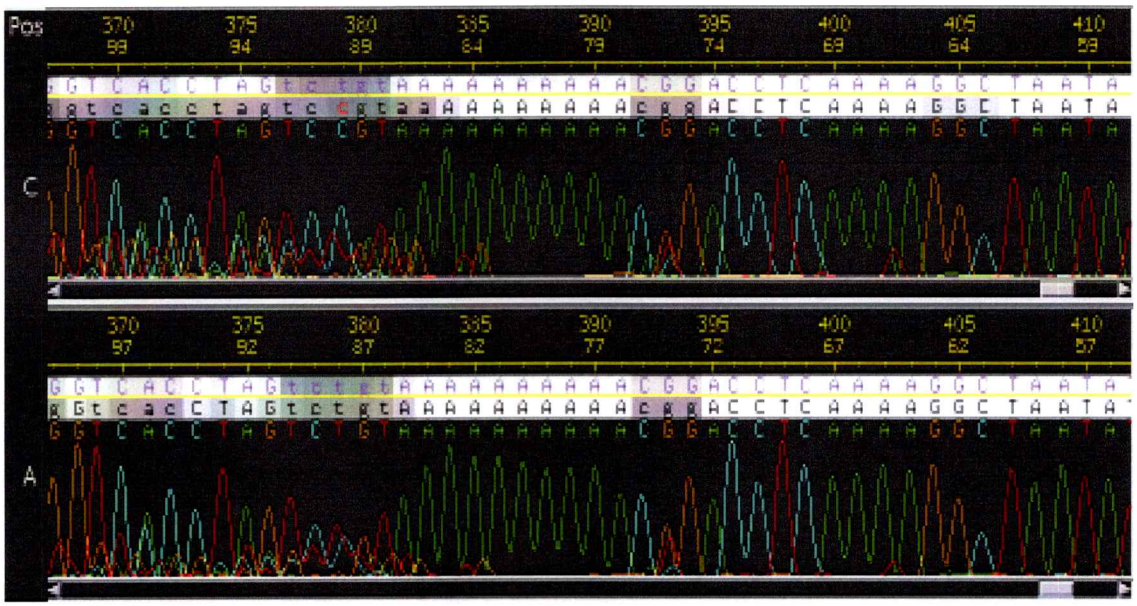

Figura B.3: **Cromatograma de región 8.** La figura muestra la presencia de múltiples peaks sobrelapados a cada lado de la región poly-N. Este hecho se puede producir por el salto de la polimerasa explicado en el texto o por la presencia de dos templados (regiones parálogas).

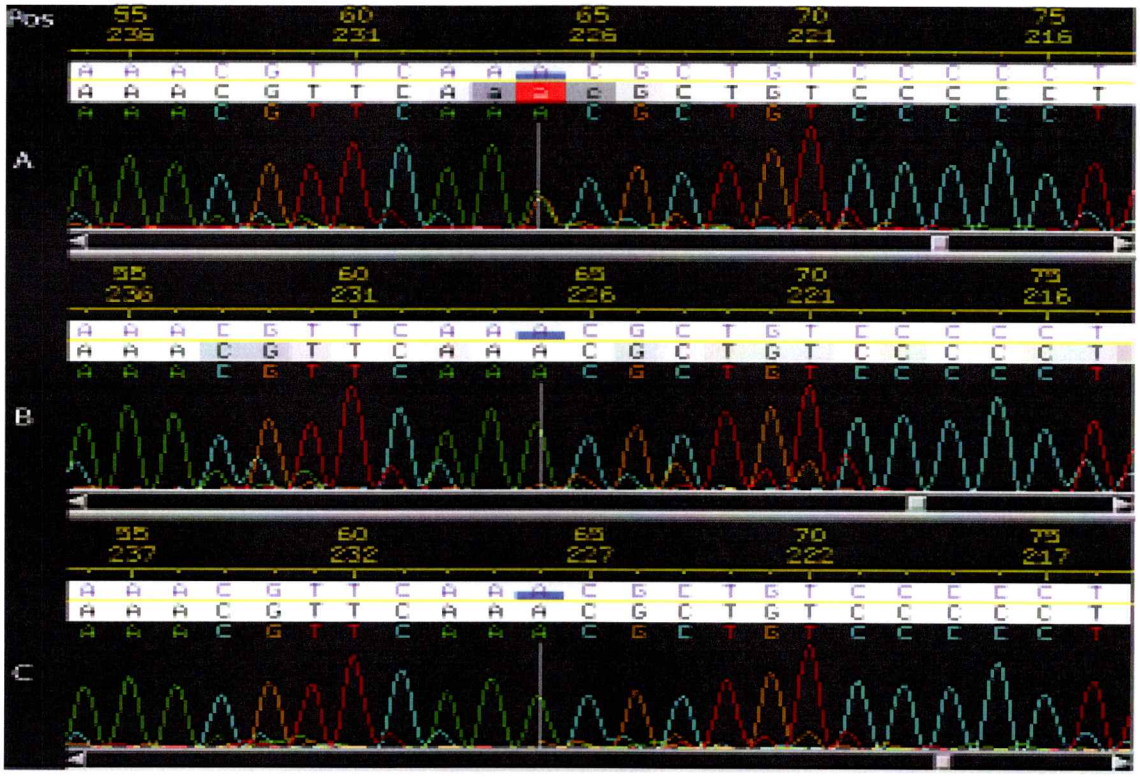

Figura B.4: **Cromatograma de la región 10.** En la figura se aprecia la presencia de heterocigotos para el SNP buscado.

C. Glosario y lista de abreviaturas

BAC : Bacterial Artificial Chromosome, librería de clones usando bacterias.

Baja complejidad : Regiones que presentan secuencias cortas repetidas sin un patrón claro ni preciso.

BES : BAC End Sequences, secuencias provenientes de los extremos de clones BAC.

BioPerl : Proyecto bioinformático de codigo abierto que permite desarrollar scripts para el análisis de datos genómicos, entre otros.

CDS : Coding sequence, secuencia codificante.

cGRASP : Genome Research in Atlantic Salmon Project Consortium.

CNVs : Copy Number Variation, variación en el número de copias.

Cobertura : Número de lecturas de DNA que dan cuenta de la misma posición nucleotídica a lo largo de una secuencia.

Contigs : Agrupación de más de una secuencias con similitud determinada por el ensamble.

Cromatograma : También llamado electroferograma. Corresponde al gráfico que muestra la secuencia de datos producida por una mÃ¡quina de secuenciación de DNA.

Deriva Génica : Se refiere al cambio estocástico en la frecuencia poblacional de una mutación determinada dada por el proceso de muestreo inherente de la reproducción.

DDC : Duplication, Degeneration, Complementation process, modelo de preservación de duplicados por duplicación, degeneración y complementación.

EAC : Escape from Adaptive Conflict, modelo de preservación de duplicados por escape del conflicto adaptativo (ver introducción para referencias).

Elementos Repetitivos : Secuencias de DNA de largo arbitrario repetidas múltiples veces a lo largo del genoma.

EST : Expressed Sequence Tag, secuencias expresadas provenientes de alguna biblioteca de cDNA.

E-value : Valor esperado en el numero de alineamientos posibles que ocurrirían al azar de una secuencia determinada contra una base de datos específica. A menor valor mas significativo el resultado.

Enmsacarar : Término asociado al proceso de identificar cierto tipo de secuencia y excluirla de los análisis posteriores, sin perder el contenido de esta.

Ensamblaje : Proceso inform´Ã¡tico en el que se comparan múltiples secuencias de DNA y se evalúa, bajo ciertos parámetros, si pertenecen a un mismo grupo o contig.

Fijación : Situación en el que una mutación ha alcanzado una frecuencia de 100 % en una población natural.

FISH : Fluorescent In Situ Hibridization, técnica de hibridación *in situ* por sondas fluorescentes.

GHM : Gene Homology Matrix, matriz para la búsqueda de homología entre genes.

Hit Blast : Resultado entregado por el programa de alineamiento Blast.

Homólogo : En genética se refiere a un par de secuencias que comparten funciones semejantes y se derivan de un origen común.

Indels : Inserción ó duplicación.

Ks : Tasa de sustituciones sinónimas por sitio sinónimo.

Ka : Tasa de sustituciones no sinónimas por sitio no sinónimo.

LINEs : Long Interspersed Elements, elementos transponibles clase I de secuencia larga.

LTRs : Long Terminal Repeats. Repeticiones de secuencias corta ubicadas en los extremos de secuencias transponibles.

MAF : Minor Allele Frequency, frecuencia de la variante alélica en menor proporción.

Mapa físico : Mapa generado por el patrón de restricción dado en una librería genómica.

MAS : Marker Assisted Selection, selección asistida por marcadores.

MSV : Multiple Sequence Variants, variantes producto de sitios polimórficos en copias duplicadas.

NF : Número de brazos cromosómicos.

Ortólogo : El término se refiere a genes que cumplen la misma función en diferentes especies.

Parálogo : El término se refiere a genes que no necesariamente cumplen la misma función pero tienen un origen común dado por un evento de duplicación.

Presión selectiva : Se refiere a cualquier fenómeno externo que altera la capacidad de supervivencia de un organismo vivo bajo un ambiente determinado.

Porcentaje de variación : Proporción de lecturas que dan cuenta de un alelo con respecto al total de lecturas presentes en una posición nucleotídica específica.

Pseudogen : Secuencia nucleotídica similar a un gen normal pero que no da como resultado un producto funcional.

PSV : Paralogous Sequence Variants, variantes producto de diferencias en la secuencia de copias parálogas.

RFLP : Restriction Fragment Length Polymorphism, polimorfismos producto de diferencias observadas en el patrón de restricción.

SINEs : Short Interspersed transposable Elements, elementos transponibles clase I de secuencia corta.

Sustitución sinónima : Cambio de base en una secuencia codificante que no altera el aminoácido codificado por el codón respectivo.

Sustitución no sinonima : Cambio de base en una secuencia codificante que altera el aminoácido codificado por el codón respectivo.

Script : Conjunto de instrucciones utilizando lenguaje informático.

Secuencia consenso : Secuencia de DNA única y de referencia generada por el alineamiento de múltiples lecturas provenientes de una misma región.

Selección natural : Mecanismo evolutivo que se define como la forma en que el ambiente externo a un organismo, dentro de una población, afecta la reproducción exitosa de éste a partir de características determinadas.

Selección Positiva : Selección actuante sobre mutaciones que le confieren una ventaja adaptativa al organismo.

Selección Negativa : Selección encargada de eliminar de la población las mutaciones deletéreas.

Selección Purificadora : Selección encargada de eliminar de la población las mutaciones deletéreas o que provoquen un cambio directo en la función del gen.

Singletons : Contigs con una lectura.

SNPs : Single Nucleotide Polymorphisms, variantes de secuencia en un sólo nucelótido.

SSCP : Single Strand Conformation Polymorphism, polimorfismos producto de diferencias observadas en su estructura secundaria en gel denaturante.

SSR : Short Sequence Repeats ó microsatélites.

Suite : Conjunto de programas informáticos bajo un mismo nombre representativo.

Tandem : Término asociado a una serie de elementos repetidos consecutivamente.

Touchdown PCR : Protocolo de PCR que incorpora una disminución de la temperatura de annealing cada cierto número de ciclos.

UTR : Untranslated Region, regiones no traducidas de los genes.

WGD : Whole genome duplication, duplicación de genoma completo.

YAC : Yeast Articial Chromosome, libreriía de clones usando levadura.

Bibliografía

ADZHUBEI, A.A., VLASOVA, A.V., HAGEN-LARSEN, H., RUDEN, T. A., LAERDAHL, J. K., HØYHEIM, B., 2007. Annotated expressed sequence tags (ESTs) from pre-smolt Atlantic salmon (Salmo salar) in a searchable data resource. BMC Genomics, **8**:209.

ALLENDORF, F.W., THORGAARD, G.H., 1984. Tetraploidy, the evolution of salmonoid fishes. Turner, J.B. (Ed.) Evolutionary Genetics of Fishes.:1–53.

AMORES, A., FORCE, A., YAN, Y.L., JOLY, L., AMEMIYA, C., FRITZ, A., HO, R.K., LANGELAND, J., PRINCE, V.E., WANG, Y.L., 1998. Zebrafish hox clusters, vertebrate genome evolution. Science, **282**:1711.

ARANEDA, C., NEIRA, R., LAM, N., ITURRA, P., 2008. Genome Mapping, Genomics in Fishes, Acquatic Animals, vol. 2. Springer-Verlag Berlin Heidelberg.

BAER, C.F., MIYAMOTO, M.M., DENVER, D.R., 2007. Mutation rate variation in multicellular eukaryotes: causes, consequences. Nat. Rev. Genet., **8**:619–631.

BAILEY, G.S., POULTER, R.T., STOCKWELL, P.A., 1978. Gene duplication in tetraploid fish: Model for gene silencing at unlinked duplicated loci. Proc. Natl. Acad. Sc., **75**:5575–5579.

BENSON, G., 1999. Tandem repeat finder: A program to analyze DNA sequences. Nucleic Acids Res, **27**:573–580.

BIRNEY, E., COPLEY, R., 2001. Wise2: intelligent algorithms for DNA searches.

BOULDING, E.G., CULLING, M., GLEBE, B., BERG, P.R., LIEN, S., MOEN, T., 2008. Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. Heredity, **101**:381–391.

BUETOW, K.H., EDMONSON, M.N., CASSIDY, A.B., 1999. Reliable identification of large numbers of candidate SNPs from public EST data. Nature genet., **21**:323–325.

CHAIN, F.J.J., ILIEVA, D., EVANS, B.J., 2008. Duplicate gene evolution, expression in the wake of vertebrate allopolyploidization. BMC Evol. Biol., **8**:43.

CONANT, G.C., WAGNER, A., 2003. Asymmetric sequence divergence of duplicate genes. Genome Res., **13**:2052–2058.

CONANT, G.C., WOLFE, K.H., 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet., **9**:938–950.

DANZMANN, R.G., DAVIDSON, E.A., FERGUSON, M.M., GHARBI, K., KOOP, B.F., HØYHEIM, B., LIEN, S., LUBIENIECKI, K.P., MOGHADAM, H.K., PARK, J., OTHERS., 2008. Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout, Atlantic salmon). BMC Genomics, **9**:557.

DAVIDSON, W.S., KOOP, B., 2008. Genomics, the Genome Duplication in Salmonids. En TSUKAMOTO, K., Y COL. (eds.), Fisheries for global Welfare and Environment, TERRAPUB 2008, pp. 77–86.

DE BOER, J.G., YAZAWA, R., DAVIDSON, W.S., KOOP, B.F., 2007. Bursts, horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics, **8**:422.

DES MARAIS, D.L., RAUSHER, M.D., 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature, **454**:762–765.

ENGELS, R., YU, T., BURGE, C., MESIROV, J.P., DECAPRIO, D., GALAGAN, J.E., 2006. Combo: a whole genome comparative browser. Bioinformatics, **22(14)**:1782–3.

EWING, B., HILLIER, L., WENDL, M.C., GREEN, P., 1998. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. Genome Res., **8**:175–185.

FESCHOTTE, C., 2008. Transposable elements, the evolution of regulatory networks. Nat. Rev. Genet., **9**:397–405.

FORCE, A., LYNCH, M., PICKETT, F.B., AMORES, A., YAN, Y., POSTLETHWAIT, J., 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics, **151**:1531–1545.

FREDMAN, D., WHITE, S.J., POTTER, S., EICHLER, E.E., DEN DUNNEN, J.T., BROOKES, A.J., 2004. Complex SNP-related sequence variation in segmental genome duplications. Nature genetics, **36(8)**:861–866.

GARZIA, L., D'ANGELO, A., AMORESANO, A., KNAUER, S.K., CIRULLI, C., CAMPANELLA, C., STAUBER, R.H., STEEGBORN, C., IOLASCON, A., ZOLLO, M., 2008. Phosphorylation of nm23-H1 by CKI induces its complex formation with h-prune, promotes cell motility. Oncogene, **27**:1853–1864.

GILBEY, J., VERSPOOR, E., MCLAY, A., HOULIHAN, D., 2004. A microsatellite linkage map for Atlantic salmon (Salmo salar). Anim. Genet., **35**:98–105.

GILLESPIE, J. H., 1994. The causes of molecular evolution. Oxford University Press, USA.

GOFFAUX, F., CHINA, B., DAMS, L., CLINQUART, A., DAUBE, G., 2005. Development of a genetic traceability test in pig based on single nucleotide polymorphism detection. Forensic Sci. Int., **151**:239–247.

GORDON, D., ABAIJAN, C., GREEN, P., 1999. Consed: A graphical tool for sequences finishing. Genome Research, **8**:195–202.

GRANDJEAN, F., VERNE, S., CHERBONNEL, C., RICHARD, A., 2009. Fine-scale genetic structure of Atlantic salmon using microsatellite markers: effects of restocking, natural recolonization. Freshwater Biology, **54**:417.

GRIFFITHS, A. J. F., WESSLER, S. R., LEWONTIN, R. C., GELBART, W. M., SUZUKI, D. T., MILLER, J. H., 2005. Introduction to genetic analysis. WH Freeman, 8th ed.

GUT, I.G., LATHROP, G.M., 2004. Duplicating SNPS. Nat. Genet., **36**:789–790.

HAN, M.V., DEMUTH, J.P., MCGRATH, C.L., CASOLA, C., HAHN, M.W., 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res., **19**:859–867.

HASTEIN, T., BERTHE, F., HILL, B.J., 2001. Traceability of aquatic animals. Rev. sei. teeh. Off int. Epiz., **20**:564–583.

HASTINGS, P. J., LUPSKI, J.R., ROSENBERG, S.M., IRA, G., 2009. Mechanisms of change in gene copy number. Nature Reviews Genetics, **10**:551.

HAYES, B., SONESSON, A., GJERDE, B., 2005. Evaluation of three strategies using DNA markers for traceability in aquaculture species. Aquaculture, **250**:70.

HAYES, B., HE, J., MOEN, T., BENNEWITZ, J., 2006. Use of molecular markers to maximise diversity of founder populations for aquaculture breeding programs. Aquaculture, **255**:573.

HAYES, B., LAERDAHL, J.K., LIEN, S., MOEN, T., BERG, P.R., HINDAR, K., DAVIDSON, W.S., KOOP, B.F., ADZHUBEI, A.A., HØYHEIM, B., 2007a. An extensive resource of single nucleotide

polymorphism markers associated with Atlantic salmon (Salmo salar) expressed sequences. Aquaculture, **265**:82.

HAYES, B., NILSEN, K.J., BERG, P.R., GRINDFLEK, E., LIEN, S., 2007b. SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. Bioinformatics, **23**:1692–1693.

HE, X., ZHANG, J., 2005. Rapid subfunctionalization accompanied by prolonged, substantial neofunctionalization in duplicate gene evolution. Genetics, **169**:1157–1164.

HOEGG, S., BRINKMANN, H., TAYLOR, J.S., MEYER, A., 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J. Mol. Evol., **59**:190–203.

HUANG, X., MADAN, A., 1999. CAP3: A DNA Sequence Assembly Program. Genome Research, **9**:868.

HUFTON, A.L., GROTH, D., VINGRON, M., LEHRACH, H., POUSTKA, A.J., PANOPOULOU, G., 2008. Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. Genome Res., **18**:1582–1591.

INNAN, H., KONDRASHOV, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics, **11(2)**:97–108.

KASAHARA, M., 2007. The 2R hypothesis: an update. Curr. Opin. Immunol., **19**:547–552.

KAZAZIAN, H.H., 2004. Mobile elements: drivers of genome evolution. Science, **303**:1626–1632.

KIM, S., MISRA, A., 2007. SNP genotyping: technologies and biomedical applications. Annual review of biomedical engineering, **9**:289.

KIMURA, M., 1983. The neutral theory of molecular evolution. Cambridge University Press.

KONDRASHOV, F.A., KOONIN, E.V., 2001. Origin of alternative splicing by tandem exon duplication. Human Molecular Genetics, **10**:2661–2669.

KRATZ, E., DUGAS, J.C., NGAI, J., 2002. Odorant receptor gene regulation: implications from genomic organization. Trends in Genetics, **18**:29–34.

LI, W.H., GRAUR, D., 2000. Fundamentals of molecular evolution. Sinauer Associates Sunderland, MA.

MACCRIMMON, H. R., GOTS, B. L., 1979. World distribution of Atlantic salmon, Salmo salar. Journal of the Fisheries Research Board of Canada, **36**:422–457.

MAKI, H., 2002. Origins of spontaneous mutations: specificity, directionality of base-substitution, frameshift,, sequence-substitution mutageneses. Annu. Rev. Genet., **36**:279–303.

MCGINNIS, S., MADDEN, T.L., 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research, **32**:20.

MEYER, A., VAN DE PEER, Y., 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays, **27**:937–945.

MITCHELL, L., 2004. Fates of duplicated regions of the Atlantic Salmon Genome. Master's thesis, Deparment of Molecular Biology, Biochemistry, Simon Fraser University, Burnaby, British Columbia.

MOEN, T., FJALESTAD, K.T., MUNCK, H., GOMEZ-RAYA, L., 2004. A multistage testing strategy for detection of quantitative trait Loci affecting disease resistance in Atlantic salmon. Genetics, **167**:851–858.

MOEN, T., HAYES, B., BARANSKI, M., BERG, P.R., KJOGLUM, S., KOOP, B.F., DAVIDSON, W.S., OMHOLT, S.W., LIEN, S., 2008. A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. BMC Genomics, **9**:223.

NADEAU, J.H., SANKOFF, D., 1997. Comparable Rates of Gene loss, Functional Divergence After

Genome Duplications Early in Vertebrate Evolution. Genetics, **147**:1259–1266.

NELSON, J. S., 2006. Fishes of the world.

NG, S.H., ARTIERI, C.G., BOSDET, I.E., CHIU, R., DANZMANN, R.G., DAVIDSON, W.S., FERGUSON, M.M., FJELL, C. D., HØYHEIM, B., JONES, S. J., DE JONG, P.J., KOOP, B.F KRZYWINSKI M.I., LUBIENIECKI, K., MARRA, M.A., MITCHELL, L.A., MATHEWSON, C., OSOEGAWA, K., PARISOTTO, S.E., PHILLIPS, R., RISE, M.L., VON SCHALBURG, K.R., SCHEIN, J. E., SHIN, H., SIDDIQUI, A., THORSEN, J., WYE, N., YANG, G., ZHU, B., 2005. A physical map of the genome of Atlantic salmon, Salmo salar. Genomics, **86**:396–404.

NICHOLS, K.M., YOUNG, W.P., DANZMANN, R.G., ROBISON, B. D., REXROAD, C., NOAKES, M., PHILLIPS, R., BENTZEN, P., SPIES, I., KNUDSEN, K., 2003. A consolidated linkage map for rainbow trout (Oncorhynchus mykiss). Animal Genetics, **34**:102.

NICKERSON, D. A., TOBE, V. O., TAYLOR, S. L., 1997. PolyPhred: automating the detection, genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Research, **25**:2745.

NIELSEN, R., 2005. Molecular Signatures of Natural Selection. Annnu. Rev. Genet., **39**:197–218.

NIELSEN, R., HELLMANN, I., HUBISZ, M., BUSTAMANTE, C., CLARK, A.G., 2007. Recent, ongoing selection in the human genome. Nat. Rev. Genet., **8**:857–868.

NOVAK, A. E., JOST, M. C., LU, Y., TAYLOR, A. D., ZAKON, H. H., RIBERA, A. B., 2006. Gene duplications, evolution of vertebrate voltage-gated sodium channels. Journal of molecular evolution, **63**:208–221.

OHNO, S., 1970. Evolution by Gene Duplication. Springer-Verlag, New york - Heidelberg - Berlin.

OJIMA, N., YAMASHITA, M., WATABE, S., 2005. Comparative expression analysis of two paralogous Hsp70s in rainbow trout cells exposed to heat stress. Biochim. Biophys. Acta, **1681**:99–106.

PALTI, Y., GAHR, S. A., HANSEN, J. D., REXROAD, C. E., 2004. Characterization of a new BAC library for rainbow trout: evidence for multi-locus duplication. Anim. Genet., **35**:130–133.

PHILLIPS, R., RAB, P., 2001. Chromosome evolution in the Salmonidae (Pisces) : an update. Biol. Rev., **76**:1–25.

POP, M., PHILLIPPY, A., DELCHER, A.L., SALZBERG, S.L., 2004. Comparative genome assembly. Briefings in Bioinformatics, **5**:237–248.

POSADA, D., 2003. Using Modeltest, PAUP to select a model of nucleotide substitution. Current Protocols in Bioinformatics.

PRINCE, V.E., PICKETT, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet., **3**:827–837.

RASTOGI, S., LIBERLES, D. A., 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol., **5**:28.

RENGMARK, A., 2006. Genetic variability in wild, farmed Atlantic salmon (Salmo salar) strains estimated by SNP, microsatellites. Aquaculture, **253**:229.

RISE, M.L., VON SCHALBURG, K.R., BROWN, G.D., MAWER, M.A., DEVLIN, R.H., KUIPERS, N., BUSBY, M., BEETZ-SARGENT, M., ALBERTO, R., GIBBS, A.R., HUNT, P., SHUKIN, R., ZEZNIK, J.A., JONES, S.R.M., NELSON, C., SMAILUS, D.E., JONES, S.J.M., SCHEIN, J.E., MARRA, M.A., BUTTERFIELD, Y.S.N., STOTT, J.M., NG, S.H., DAVIDSON, W.S., KOOP, B.F., 2004. Development, application of a salmonid EST database, cDNA microarray: data mining, interspecific hybridization characteristics. Genome Res., **14**:478–490.

ROBINSON-RECHAVI, M., BOUSSAU, B., LAUDET, V., 2004. Phylogenetic dating, characterization of gene duplications in vertebrates: the cartilaginous fish reference. Mol. Biol. Evol., **21**:580–

586.
ROZEN, S., SKALETSKY, H.J., 2000. Primer 3 on the WWW for general users, biologist programmers., vol. 132. Humana Pr Inc, pp. 365–386.
RYYNANEN, H.J., PRIMMER, C.R., 2006. Single nucleotide polymorphism (SNP) discovery in duplicated genomes: intron-primed exon-crossing (IPEC) as a strategy for avoiding amplification of duplicated loci in Atlantic salmon (Salmo salar) and other salmonid fishes. BMC Genomics, **7**:192.
RYYNANEN, H.J., TONTERI, A., VASEMAGI, A., PRIMMER, C.R., 2007. A comparison of biallelic markers, microsatellites for the estimation of population, conservation genetic parameters in Atlantic salmon (Salmo salar). J. Hered., **98**:692–704.
SEMON, M., WOLFE, K.H., 2007. Consequences of genome duplication. Curr. Opin. Genet. Dev., **17**:505–512.
SMIT, A.F.A., HUBLEY, R., GREEN, P., 1996. RepeatMasker Open-3.0., http://www.repeatmasker.org.
SMITH, C.T., 2005. Characterization of 13 single nucleotide polymorphism markers for chum salmon. Molecular Ecology Notes, **5**:259.
SMITH, C.T., ELFSTROM, C.M., SEEB, L.W., SEEB, J.E., 2005. Use of sequence data from rainbow trout, Atlantic salmon for SNP detection in Pacific salmon. Mol. Ecol., **14**:4193–4203.
TAYLOR, J.S., RAES, J., 2004. Duplication, divergence: the evolution of new genes, old ideas. Annu. Rev. Genet., **38**:615–643.
THOMPSON, J.D., HIGGINS, D.G., GIBSON, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, **22(22)**:4673.
THORGAARD, G. H., BAILEY, G.S., WILLIAMS, D., BUHLER, D. R., KAATTARI, S. L., RISTOW, S. S., HANSEN, J. D., WINTON, J. R., BARTHOLOMEW, J. L., NAGLER, J. J., WALSH, P.J., VIJAYAN, M.M., DEVLIN, R.H., HARDY, R.W., OVERTURF, K.E., YOUNG, W.P., ROBISON, B. D., REXROAD, C., PALTI, Y., 2002. Status, opportunities for genomics research with rainbow trout. Comparative Biochemistry, Physiology, Part B, **133**:609–646.
THORSEN, J., ZHU, B., FRENGEN, E., OSOEGAWA, K., DE JONG, P.J., KOOP, B. F., DAVIDSON, W.S., HØYHEIM, B., 2005. A highly redundant BAC library of Atlantic salmon (Salmo salar): an important tool for salmon projects. BMC Genomics, **6**:50.
TSUCHIHASHI, Z., DRACOPOLI, N.C., 2002. Progress in highthroughput SNP genotyping. Pharmacogenomics Journal, **2**:103–110.
UTTER, F.M., ALLENDORF, F.W., HODGINS, H.O., 1973. Genetic variability, relationships in Pacific Salmon, Related Trout based on protein variations. Systematic Zoology, **22**:257–270.
VANDEPOELE, K., DE VOS, W., TAYLOR, J. S., MEYER, A., VAN DE PEER, Y., 2004. Major events in the genome evolution of vertebrates: paranome age, size differ considerably between ray-finned fishes, land vertebrates. Proceedings of the National Academy of Sciences, **101**:1638–1643.
VIGNAL, A., MILAN, D., SANCRISTOBAL, M., EGGEN, A., 2002. A review on SNP, other types of molecular markers, their use in animal genetics. Genet. Sel. Evol., **34**:275–305.
VOLFF, J.N., 2005. Genome evolution, biodiversity in teleost fish. Heredity, **94**:280–294.
WEBB, J.H., VERSPOOR, E., AUBIN-HORTH, N., ROMAKKANIEMI, A., AMIRO, P., 2006. The Atlantic Salmon., chap. 2. Blackwell Publishing, Oxford, pp. 17–56.
WILLSON, M. F., 1997. Variation in Salmonid Life: Patterns, perspectives.
WOLFE, K.H., SEOIGHE, C., 1999. Updated map of duplicated regions in the yeast genome. Gene

238:253–261.

XU, P., WANG, S., LIU, L., PEATMAN, E., SOMRIDHIVEJ, B., THIMMAPURAM, J., GONG, G., LIU, Z., 2006. Channel catfish BAC-end sequences for marker development, assessment of syntenic conservation with other fish species. Anim. Genet., **37**:321–326.

YANG, Z.H., BIELAWSKI, J.P., 2000. Statistical methods for detecting molecular adaptation. Trends. Ecol. Evol, **15**:496–503.

YURYEV, A., 2007. PCR primer design. Humana Pr Inc.

ZHANG, J., 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution, **18**:292–298.

ZHANG, Z., LI, J., ZHAO, X.Q., WONG, G.K., YU, J., 2006. KaKs Calculator: Calculating Ka, Ks Through Model Selection, Model Averaging. Geno. Prot. Bioinfo., **4**:259–263.

ZHAO, S., MALEK, J., MAHAIRAS, G., FU, L., NIERMAN, W., ADAMS, M.D., VENTER, J.C., 2000. Human BAC ends quality assessment, sequence analyses. Genomics, **63**:321–332.

ZHAO, S., SHATSMAN, S., AYODEJI, B., GEER, K., TSEGAYE, G., KROL, M., GEBREGEORGIS, E., SHVARTSBEYN, A., RUSSELL, D., OVERTON, L., JIANG, L., DIMITROV, G., TRAN, K., SHETTY, J., MALEK, J.A., FELDBLYUM, T., NIERMAN, W.C., FRASER, C.M., 2001. Mouse BAC ends quality assessment, sequence analyses. Genome Res., **11**:1736–1745.

ZHAO, Z., FU, Y.X., HEWETT-EMMETT, D., BOERWINKLE, E., 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome, its implications for molecular evolution. Gene, **312**:207–213.