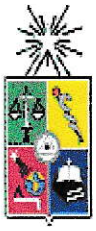




# VALIDACIÓN DE SNPS PARA EL DESARROLLO DE ESTRATEGIAS DE ESTUDIOS DE ASOCIACIÓN GENÓMICA, EN SALMONES.

Tesis

Entregada a la

Universidad de Chile

En cumplimiento parcial de los requisitos

Para optar al Grado de

Magíster en Ciencias con Mención en Biología Molecular, Celular y Neurociencias

Facultad de Ciencias

Por

Guillermo Antonio Rodríguez Piccoli

Director de Tesis: Dr. Miguel Allende

Junio, 2013

Santiago – Chile

ESCUELA DE POSTGRADO – FACULTAD DE CIENCIAS – UNIVERSIDAD DE CHILE

# INFORME DE APROBACIÓN

# TESIS DE MAGISTER

Se informa a la Escuela de Postgrado de la Facultad de Ciencias que la Tesis de Magíster presentada por el candidato:

Sr. GUILLERMO ANTONIO RODRÍGUEZ PICCOLI

Ha sido aprobada por la comisión de Evaluación de la tesis como requisito para optar al grado de Magíster en Ciencias con mención en Biología Molecular, Celular y Neurociencias, en el examen de Defensa Privada de Tesis rendido el día:

25 de Abril del 2013

**Director de Tesis**

*Dr. Miguel Allende*

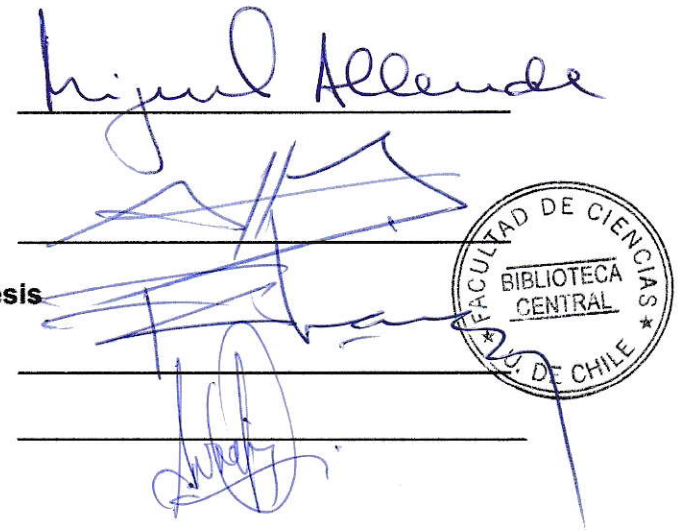
**Co-Director de Tesis**

*Dr. Alejandro Maass*

**Comisión de Evaluación de la Tesis**

*Dr. Francisco Chávez*

*Dr. Francisco Perez*



# DEDICATORIA

Quisiera dedicar el presente trabajo de tesis a mi familia, mi madre Daniella, mi abuela Oda, a mis tíos Renzo y Claudia, a mi padre Luis, a mis hermanos y a todos quienes como ellos me han apoyado durante mi educación superior y en la vida para llegar a ser quien soy.

# AGRADECIMIENTOS

Quisiera agradecer a mi Director de Tesis el Dr. Alejandro Maass, por guiarme en el desarrollo de este trabajo y en el enfoque de mi carrera profesional; al jefe de laboratorio de Mathomics Alex Di Genova Bravo por guiarme en el enfoque bioinformático del trabajo presente;a Luis Zapata Ortiz que su trabajo de Tesis de pregrado dio pie para la realización de este trabajo; ami Co-Director de Tesis el Dr. Miguel Allende Connely por su patrocinio para la realización y aprobación de esta tesis inter-facultad con enfoques industriales;a los integrantes de los laboratorios Mathomics y CRG, por su ayuda en el desarrollo del proyecto, especialmente a Andrés Aravena, Marko Budinich, María Paz Cortes, Nicolás Loira y Dante Travisany por responder pacientemente mis incesantes preguntas; ala empresa AQUAINNOVO S.A.,y el gobierno de Chile con el Fondo de Financiamiento de Centros de excelencia en Investigación (FONDAP) y las becas de la Comisión Nacional de Investigación Científica y Tecnológica de Chile(CONICYT),que hicieron posible financiar el presente proyecto de investigación; y finalmente a Ben Koop de cGRASP por facilitarme las bases de datos de elementos repetitivosya ISCASG por el ensamble utilizado como referencia en el presente trabajo.

# INDICE DE MATERIAS

# LISTA DE TABLAS

# LISTA DE FIGURAS

# LISTA DE ABREVIATURAS

**[1|2|3|4]R-WGD (*[1|2|3|4]Round of Whole Genome Duplication*)**, ronda de duplicación de genoma completo número [1|2|3|4].

**BAC (*Bacterial Artificial Chromosome*)**, constructo de DNA basado en un plásmido de fertilidad funcional, utilizado para transformar y clonar DNA en bacteria.

**BAM(*Binary Sequence Alignment/Map*)**, versión binaria comprimida del formato de alineamiento de secuencia SAM.

**BAQ (*Base Alignment Quality*)**, probabilidad Phred de que una base esté mal alineada.

**BES(*BAC-End Sequencing*)**, secuenciamiento de las regiones terminales de secuencias pertenecientes a una librería de clones BAC.

**BWT (*Burrows-Wheeler Transform*)**,algoritmo de compresión utilizado en NGS para reducir los requerimientos de memoria del alineamiento de secuencias.

**cDNA(*complementary DNA*)**, DNA complementario sintetizado a partir de un templado de mRNA, catalizado por transcriptasa reversa y DNA polimerasa.

**CDS(*Coding DNA Sequence*)**, región codificante de un gen compuesta por zonas exónicas.

**CI(*Confidence Interval*)**, indicador estadístico de la confianza de un intervalo dentro de un parámetro poblacional.

**CMM** (***C**enter for **M**athematical **M**odeling*), centro de investigación de modelamiento matemático de la Universidad de Chile.

**COMAV**, Instituto Universitario de **CO**nservación y **M**ejora de la **A**grodiversidad Valenciana.

**CONICYT**, **CO**misión **N**acional de Investigación **C**ientífica **YT**ecnológica, Gobierno de Chile.

**DIP**(*Deletion/Insertion Polymorphism*), Secuencia corta de nucleótidos presente en un sub-grupo de miembros de la población.

**DNA**(*DeoxyRibonucleicAcid*), ácido nucleico que contiene las instrucciones genéticas utilizadas en el desarrollo y funcionamiento de los seres vivos conocidos.

**RE** (*Repetitive Element*), secuencias de DNA repetidas a lo largo del genoma. Pueden ser de dos tipos, repetidos en tándem y repeticiones intercaladas.

**EST**(*Expressed Sequence Tag*), corta subsecuencia de una hebra de cDNA.

**FAO**(*Food and Agriculture Organization of the United Nations*), foro neutral donde todas las naciones participantes se juntan por igual a negociar acuerdos y políticas de debate.

**FL**(*Fuzzy Logic*), forma de lógica de valores múltiples o lógica probabilística; trata el razonamiento lógico que es aproximado en lugar de fijo y exacto.

**FONDAP**, **FON**do **D**e Investigación Avanzado en **Á**reas **P**rioritarias para Centros de Excelencia.

**GATK(*Genome Analysis ToolKit*)**, paquete de software desarrollado por"Broad Institute" para el análisis de datos de NGS.

**GB (*GigaBytes*)**, unidad de información digital equivalente a diez a la nueve bytes, generalmente aceptados como equivalentes a 8 bits (pulsos de voltaje).

**Gb (*Gigabases*)**, unidad de información génica equivalente a diez a la nueve bases nucleotídicas.

**gDNA(*genomic DNA*)**, DNA cromosomal consecutivo, diferente de cDNA o plásmidos.

**GEBV(*Genomic Breeding Value Estimator*)**, estimador del valor genómico de cruza de dos individuos de un ganado con respecto a un conjunto determinado de rasgos de interés.

**GHz (*GigaHertz*)**, Unidad de frecuencia de transmisión de información por pulsos de voltaje (bits) equivalente a diez a la nueve Hertz (Hz).

**GMAP(*Genomic Mapping andAlignment Program*)**, programa bioinformático para el mapeo y alineamiento de secuencias con gaps (i.e. mRNA e EST) a un genoma de referencia; basado en algoritmo de hashing.

**GNU (*GNU's NotUnix*)**, sistema operativo basado en Unix desarrollado por el proyecto GNU.

**GWAS(*Genome-WideAssociation Study*)**, estudio de asociación de genoma completo.

**HPCLAB (*High Performance ComputingLaboratory*)**, laboratorio de computación de alto rendimiento del CMM.

**Hz (*Hertz*)**, unidad de frecuencia equivalente a un evento por segundo.

**ICSASG** (*InternationalCooperation to Sequence the AtlanticSalmonGenome*), cooperación internacional para el secuenciamiento del genoma del salmón del Atlántico, formada por"Genome BC", "Agencia de Desarrollo Económico de Chile","InnovaChile","Norwegian Research Council" y"Norwegian Fishery and Aquaculture Industry Research Fund".

**IGV**(*Integrative Genomics Viewer*), herramienta de visualización de alto rendimiento para la exploración interactiva de grandes conjuntos de datos genómicos integrados.

**IVs** (*Indivudual Variants*), variantes de secuencias entre un único individuo, comprendiendo las variantes entre regiones parálogas como PSVs o MSVs.

**LD**(*Linkage Disequilibrium*), asociación no aleatoria de alelos en dos o más *loci*, pertenecientes a un mismo o distinto cromosoma.

**MAS**(*Marker Assisted Selection*), proceso por el cual un marcador es utilizado para la selección indirecta de uno o varios determinantes genéticos de un rasgo de interés.

**MIVs** (*Multiple Individuals Variants*), variantes de secuencia obtenidas por mapeo y comparación de secuencias entre múltiples individuos, estas comprenden SNPs, DIPs y MSVs; así como también posibles PSVs.

**MSV**(*MultiSite Variant*), variante de un nucleótido con características complejas debida a una variación en el número de copias o conversión genómica.

**NA**(*Not Applicable*), abreviación que indica que cierto indicador no es aplicable debido a la falta de datos.

**NCBI(*NationalCenter for Biotechnology Information*)**, recurso de información biológica molecular financiado por el gobierno de estados unidos.

**NGS(*Next Generation Sequencing*)**, secuenciación de alto rendimiento son tecnologías que paralelizan el proceso de secuenciación, produciendo miles de millones de secuencias a la vez.

**PES (*Paired-End Sequencing*)**, tecnología de NGS donde se amplifican y luego secuencian fragmentos de los extremos 5' y 3' de una secuencia de DNA, con un rango de separación variable conocido, útil en scaffolding.

**PCA (*PrincipalComponent Analysis*)**, procedimiento matemático que usa una transformada ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no-correlacionadas llamadas componentes principales.

**PSV(*Paralogous Sequence Variant*)**, cambio genético que no son debidos a un polimorfismo sino a la diferencia entre nucleótidos entre dos copias parálogas de una secuencia duplicada en el genoma.

**QDR(*QuadDataRate*)**, es un tipo de memoria RAM estática capaz de transportar hasta cuatro palabras de datos por cada ciclo de frecuencia.

**QTL(*Quantitative TraitLocus*)**, conjunto de efectos poligénicos determinantes de un fenotipo dado.

**RAM(*RandomAccessMemory*)**, forma de almacenamiento de datos.

**SAM(*Sequence Alignment/Map*)**, formato genérico para el almacenamiento de grandes alineamientos de secuencias nucleotídicas.

**SNP (*Single-Nucleotide Polymorphism*)**, variación de secuencia de DNA que ocurre cuando un único nucleótido en una cierta posición del genoma difiere entre miembros de una especie biológica o cromosomas pareados dentro de un mismo individuo.

**SNPchip**, microarreglo de alta densidad de SNPs utilizado para caracterizar el genoma de un individuo y las trazas fenotípicas características de una especie.

**snpEff**, herramienta de predicción de efecto y anotación de variantes génicas.

**SNP-SACK (*SNP SearchAndCharacterizationKit Pipeline*)**, herramienta bioinformática ideada, creada y desarrollada en este proyecto de tesis para la caracterización de variantes de acuerdo a su nivel de conservación.

**SVM(*SupportVector Machine*)**, modelo de aprendizaje artificial supervisado constituido por algoritmos de aprendizaje de análisis de datos y reconocimiento de patrones.

**TB(*TeraBytes*)**, unidad de información digital equivalente a diez a la doce bytes, generalmente aceptados como equivalentes a 8 bits (pulsos de voltaje).

**TFBS(*TranscriptionFactorBinding Site*)**, sitio de unión de factor de transcripción ubicado río arriba del gen o genes a regular.

**True-SNP**, SNP funcional utilizable en MAS, terminología utilizada para diferenciar SNP debidos a diferencias entre genomas de múltiples individuos de una especie y aquellos debidos a diferencias entre secuencias parálogas.

**UTR(*UnTranslated Region*)**, regiones no traducidas ubicadas en los extremos 5' y 3' de las hebras de mRNA.

**VCF(*Variant Calling Format*)**, formato de archivo de texto que contiene información relativa a una variante en una posición específica del genoma.

**WGD(*Whole-Genome Duplication*)**, evento evolutivo que crea un organismo con copias adicionales de la totalidad de su genoma.

# RESUMEN

El salmón del Atlántico (*Salmo salar*) constituye uno de los productos de exportación más importantes paraChile y Noruega, por lo que su estudio y mejoramiento es crucial para el desarrollo de ambas economías. Esto puede ser llevado a cabo mediante herramientas moleculares, tales como marcadores genéticos (SNPs), que han demostrado su efectividad en la selección genómica asistida para el mejoramiento de otras especies (Closter, Elferink, et al., 2010). Este tipo de análisis requiere de un conjunto de SNPsde alta densidad donde cada loci del genoma se encuentre en desequilibrio de ligamiento con a lo menos uno de estos(Hayes and Goddard, 2010), haciendo esencial la correcta caracterización y clasificación de cada marcador presente en el genoma. Las múltiples duplicaciones del genoma del salmón (50% de las secuencias codificantes) (Davey et al., 2001a; Utter et al., 1973),hacen necesaria la exclusión de las abundantes estructuras polimórficas presentes en regiones parálogas de su genoma (PSVs), y la caracterización de SNPs presentes en estas zonas (MSVs) para la elección de aquellos polimorfismos con alta conservación (Hardy-Weinberg principle), que constituyan marcadores eficientes para la selección genómica. Hasta el momento no existen acercamientos para resolver esta problemática *a priori* significando altos costos y tiempos asociados a la clasificación, la cual no siempre es efectiva.Es por esto que proponemos lograr esto mediante la clasificación de los polimorfismos de múltiples individuos (MIVs) de acuerdo alos niveles de conservación de las regiones en las que se encuentran.

En el presente proyecto presentamosun pipeline bioinformático innovador, acompañadode un conjunto de herramientas único, para la búsqueda y categorización de SNPs en el genoma completo (SNP Search And Characterization Kit, SNP-SACK), y su diferenciación *a priori* de polimorfismos contaminantes, para su uso como marcadores en la selección genómica.Para esto se mapearon las secuenciaspúblicas de múltiples individuos*Salmo salar*, más secuencias adicionales Illumina obtenidas por AQUAINNOVO; y todas las lecturas públicas disponibles de "Sally", contra un

ensamble públicode "Sally", obtenidodesde ncbi. Las Estructuras génicas fueron predichas medianteun acercamiento *ab initio*y los polimorfismos utilizandoGATK. Las predicciones de conservación fueron hechas utilizando SNP-SACK,mediante un modelo de lógica difusa en 2 pasos, tomando en cuenta indicadores de conservación y densidad de variantes de secuencia en intervalos móviles a lo largo de los contigs del ensamble. Obtuvimos un conjunto final de 29.916 SNPs y 3.081 DIPs además de ~5M de candidatos PSV/MSV. El conjunto final de SNPs/DIPs consistió en 1,5K exónicos, 2,6K intrónicos, 1K en UTR y 27K en regiones intergénicas. El modelo predictivo fue validado contra un conjunto de ~ 6K de marcadores públicos caracterizados y validados (Lien et al., 2011), donde se observó que sobre un 95% de los SNPs se ubicaron en regiones de alta conservación, siendo este porcentaje un 10% menor en MSVs.Esta metodología demostró ser efectiva para la búsqueda, caracterización y filtrado de marcadores, permitiendo el filtrado de potenciales PSVs y variantes de baja conservación. Los principios detrás del presente pipeline son potencialmente aplicables a múltiples organismos con estados de caracterización similares a los del salmón del Atlántico. La base de datos resultante contiene información adicional acerca de la naturaleza de cada polimorfismo, lo que permite su potencial aplicación a otros estudios.

# ABSTRACT

Atlantic salmon (*Salmo salar*) is one of the major exports of Chile and Norway, making its study and improvement crucial to the economies of both nations. This can be achieved through molecular tools such as genetic markers (SNPs), which have demonstrated their effectiveness in genome assisted selection for the improvement of other species (Closter, Elferink, et al., 2010). These sort of analysis requires a denseSNP pool, soevery loci is in linkage disequilibrium with at least one of them(Hayes and Goddard, 2010), making essential the correct classification and characterization of each SNP present in the genome. The multiple duplications of the salmon genome (50% of coding sequences) (Davey et al., 2001b; Utter et al., 1973)make necessary the exclusion of the abundant polymorphic structures found in paralogue regions of its genome (PSVs), and the characterization of SNPs present in these regions (MSV). Our approach intends to accomplish this through the classification of the polymorphisms from multiple individuals (MIVs) according to the conservation-level of the regions on which were found.

The current work represents an innovative bioinformatic pipeline along with a unique set of tools, for the search and categorization of SNPs in the entire genome (SNP Search And Characterization Kit, SNP-SACK). SNP-SACK constitutesan *a priori* approach for the exclusion of contaminant polimorphisms, for the generation of a pool of markers for its use in genomic selection.In order to accomplish this, the public *Salmo salar*sequences from multiple individuals obtained from NCBI data base, along with an Illumina sequence pool from AQUAINNOVO, and all Sally public sequences; were mapped against a public hybrid assembly of Sally, also available atNCBI.The genetic structures were predicted using an *ab initio* approach and the polymorphisms were obtained using GATK. The predictions of conservation levels were made using a two-step fuzzy logic model (part of SNP-SACK pipeline), accounting for conservation indicators and density of sequence variants in moving intervals throughout the contigs of the assembled genome. Thefinal resulting pool consisted of 29.916 SNPs and 3.081

DIPs and ~ 5M PSVs/MSVs candidates. The final pool of SNPs/DIPs consisted of 38K exonic, 46K intronic, and 46K at UTR and 242 K in unclassified regions.This methodology has proven effective for finding, characterizing and filtering SNPs/DIPs, allowing differentiation from potential PSVs. The principles behind the present SNP/DIP search and characterization pipeline (SNP-SACK) could potentially be applied to multiple other organisms with similar states of characterization to those of Atlantic salmon. The database resulting from this process also contains information about the nature of each polymorphism, potentially lending itself to utilization in further applications.

# 1 INTRODUCCIÓN

## *1.1 Mejoramiento animal para la producción.*

### *1.1.1 Tendencias del mercado mundial e implicancias sobre el futuro de la producción animal.*

La creciente demanda mundial por proteína animal ha ejercido una gran presión sobre el sector ganadero y agropecuario para mejorar su eficiencia de producción.Este fenómeno viene dado por múltiples factores entre los cuales se cuentan el crecimiento poblacional, un aumento de los ingresos a nivel mundial y la urbanización(Who and Consultation, 2003). La organización mundial de alimentos y agricultura proyecta un incremento del73% (158 Ton.) en la demanda mundial por carnes rojas para el año 2020(FAO, 2011).

En el caso de la pescala demanda mundial se ha doblado desde 1973 (Laurenti, 2008), produciendo una disminución del 90 % en la población mundial de peces en los océanos (Myers et al., 2003). Estocausa un estancamiento en la pesca por captura que se arrastra desde la década de 1990(FAO, 2012; Panetta, 2003), lo que ha derivado en un aumento en la producción de peces por cultivo llegando a constituir un 30% de la producción mundial total(Delgado et al., 2003). Lo anterior sumado a que el agua fresca se convertirá en un recurso escaso de manera incremental en los próximos 20 años (Rosegrant et al., 2002) y latendencia mundial a la globalización,hangenerado un aumento en las vías de transmisión para enfermedades, las que siguen siendo una gran restricción sobre la producción de pesca por cultivo (Subasinghe et al., 2001).

Es por esto que el mejoramiento animal mediante la cruza es vital para el incremento de la eficiencia productiva del mercado del cultivo animal, paraasí solventar las crecientes demandas del mercado, tanto en la pesca como en el sector ganadero(Hume et al., 2011).

### *1.1.2 Mejoramiento animal por selección asistida para la cruza.*

Desde los orígenes del cultivo animal (12.000 años atrás) la humanidad ha mejorado los rasgos asociados con la productividad mediante la selección asistida de parentales para la cruza, haciendo posible obtener distintas familias de ganado con adaptaciones específicas para la elaboración de ciertos productos determinados (Serpell, 1996).

Durante los últimos años los avances sustanciales en genética molecular han permitido la identificación de regiones del genoma denominadas loci (plural del latín locus "lugar") que afectan rasgos de importancia para la producción (Andersson, 2001; Hayes, 2007). Estas tecnologías posibilitan optimizar los programas de mejoramiento animal a través de la selección de alta precisión de parentales (Meuwissen et al., 2001b), mediante marcadores moleculares asociados a estos loci (Dekkers and Hospital, 2002),acelerando la ganancia genética por generación comparado con métodos tradicionales (Bagnato and Rosati, 2012; Boichard et al., 2002; Colleau et al., 2009).

La gran mayoría de los rasgos de importancia económica para sistemas de producción en ganadería y acuicultura son cuantitativos, es decir son atribuibles a múltiples genes, resultando en una distribución continua de las posibles intensidades del rasgoa lo largo de la población (Hayes, 2007). Si consideramos lo anterior y la naturaleza finita del genoma (Ewing et al., 2000), se infiere que debe existir un número finito de loci (regiones del genoma) subyacentes a la variación de

cada rasgo cuantitativo.A estos loci se les denomina Loci del Rasgo Cuantitativo o QTLs por sus siglas en inglés (Quantitative Trait Loci, QTLs), compuesto por múltiples Loci del Rasgo Cuantitativo (Quantitative Trait Locus, QTL) divididos en un pequeño grupo de genes de gran efecto y múltiples genes de efecto pequeño (Hayes et al., 2001; Shrimpton and Robertson, 1988) (verFigura 1). La búsqueda de estos loci, particularmente los de moderado a gran efecto, y su uso para la selección eficiente de rasgos animales, ha sido una fuerte área de investigación durante las últimas décadas. La determinación de estas diferencias genéticas heredables entre individuos se denomina genotipado y es una técnica ampliamente utilizada para el mejoramiento animal.

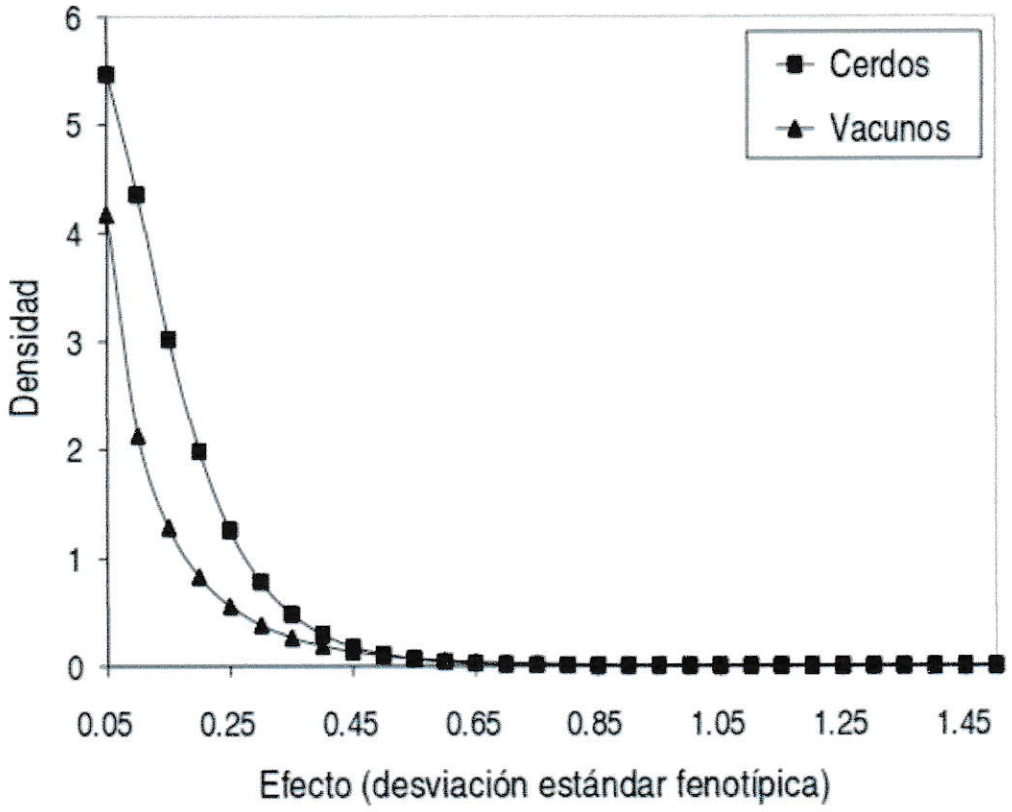



**Figura 1. Efectos de QTL de experimentos realizados en cerdo y vacuno sobre un rasgo dado ajustados a una distribución gamma por máxima verosimilitud (MLE)(Hayes, 2007).**

*El efecto del QTL se calculó como la desviación estándar fenotípica con respecto al rasgo en cuestión (Hayes, 2007), cada punto de la curva ajustada para cerdos se representa con el símbolo ■, y ▲ para vacuno.*

Se han utilizado dos aproximaciones para el descubrimiento de QTLs. La aproximación del gen candidato, asume que mutaciones en un gen involucrado en la fisiología del rasgo aportarían a la variación de éste. En esta aproximación el gen, o partes del gen candidato, son secuenciados en múltiples animales distintos, luego cualquier variación encontrada en la secuencia de DNA es testeada por asociación con la variación en el fenotipo del rasgo. Esta aproximación ha tenido algo de éxito (Andersson and Georges, 2004), e.g. se descubrió una mutación en el locus del receptor de estrógeno (ESR), que resulta en un aumento en el tamaño de camada en cerdos (Rothschild et al., 1991).Sin embargo, existen dos problemas con esta aproximación:

1) Primero, generalmente el rasgo está determinado porun gran número de genes, lo que implica un estudio de asociación para cada uno de estos genes en múltiples animales.Esto, sumado a la posibilidad de que la mutación ocurra en DNA no codificante aumenta aún más la cantidad de secuenciamiento requerido,lo que eleva considerablemente el costo y los tiempos del estudio.
2) Segundo, la o las mutaciones causantes puede que se encuentren en un gen que no sería considerado a priori como un candidato obvio para el rasgo en cuestión, resultando en una búsqueda de QTL infructuosa.

Una técnica alternativa para el descubrimiento de QTLs es la denominada"mapeo de QTL", que consiste en la identificación de loci asociados al rasgo, utilizando un conjunto de marcadores de DNA neutrales en busca de asociaciones entre las variaciones alélicas del marcador y las variaciones en el fenotipo del rasgo.Un marcador de DNA es una ubicación específica en el cromosoma cuya heredabilidad puede ser monitoreada.Estos pueden ser regiones de DNA (genes) o generalmente

algún segmento de DNA sin una función codificante conocida pero, cuyos patrones de heredabilidad pueden ser determinados (Wives and Loh, 1998).

#### *1.1.2.1 Marcadores moleculares.*

Los marcadores moleculares juegan un papel esencial en la genética animal, lo que ha llevado al desarrollo de múltiples técnicas para su obtención y una gran variedad de éstos con distintas implicancias biológicas, por lo que se deben elegir acorde al propósito del estudio a realizar.

Se deben considerar dos criterios principales al momento de utilizar los marcadores moleculares para estudios de heredabilidad. Desde el punto de vista de la biología molecular el proceso de genotipado debe ser lo más simple y económico posible, para la obtención de las enormes cantidades de datos necesarios. Desde el punto de vista estadístico y de acuerdo al tipo de análisis a realizar, se deben considerar las relaciones de dominancia, la información contenida, la neutralidad y la independencia genética de los marcadores. Esto con el objeto de generar datos tan confiables como sea posible.

Tradicionalmente, podemos clasificarlas variaciones a nivel de DNA en 3 categorías de acuerdo a su mecanismo molecular:1) cambios de un único nucleótido, a los que se les denomina SNV (Single Nucleotide Variation) si tienen una representatividad menor a un 1% dentro de una población o SNP (Single Nucleotide Polymorphsm) si su representatividad es mayor a un 1%; 2) inserciones/deleciones (oINDELs) que consisten en la existencia/falta de secuencias en ciertos individuos con respecto a la población, estos pueden tener múltiples longitudes desde 1 a cientos de pares de bases. Cuando se utiliza un INDEL como marcador se le denomina DIP (DIP, Deletion / Insertion Polymorphism),utilizándose para el genotipado generalmente solo INDELs dialélicos y de largo reducido (Weber et al., 2002; Ye et al., 2002); y 3)

variaciones en el número de secuencias repetidas en tándem o VNTR (Variations Number Tandem Repeats) también conocidas como microsatélites. Las técnicas moleculares utilizados para el genotipado se adaptan al tipo de variación y la escala y rendimiento previstos (Vignal et al., 2002).

Los marcadores de sustituciones cortas (SNPs/DIPs) presentan múltiples ventajas con respecto a otros marcadores, para la selección genómica. Dado que los DIPs tienen propiedades y utilidades similares a las de los SNPs, regularmente se les denomina genéricamente SNPs a ambos.

Un problema técnico de los microsatélites es el que dificultan la estandarización de resultados, por las eventuales inconsistencias en las determinaciones del tamaño de alelo. A pesar que esto generalmente no constituye un problema para el caso de los estudios familiares y la determinación de parentales, como los realizados por escaneo de QTLs, puede constituir un problema serio si se requiere genotipar individuos aislados, como es el caso de los estudios poblacionales. Estos problemas son causados mayoritariamente por la gran variedad de técnicas de determinación de microsatélites existentes en el mercado, cada una con diferentes marcadores fluorescentes que a su vez producen distintas migraciones en gel, entre otros factores (Hahn et al., 2001). En algunos casos estos problemas no pueden ser resueltos incluso mediante el uso de múltiples muestras estándar, particularmente cuando existen diferencias sustanciales entre los tamaños de los alelos.

Otro caso de error en la determinación del tamaño se debe a la reacción de PCR en sí: de acuerdo a las condiciones utilizadas, la Taq polimerasa cataliza la adición de una base extra (generalmente adenina) en el extremo 3' del producto de PCR. La proporción de fragmentos con esta base extra varía entre 0 a 100%, causandodiferenciasde 1 base enel largo, haciendo más complejo el análisis de los

datos. Aunque estos problemas se pueden evitar con tratamiento bioquímico luego del PCR o mediante modificación de los partidores(Brownstein et al., 1996; Ginot et al., 1996), generalmente no ocurre.

Las definiciones de alelos por microsatélites son generadas asumiendo que la variación de tamaño en los productos de PCR está directamente correlacionada con diferencias en los números de secuencias repetidas del motivo simple.

Estos problemas nombrados anteriormente y las múltiples ventajas de los marcadores de sustituciones cortas puntuales (SNPs/DIPs) para estudios que requieren paneles de marcadores de alta densidad y precisión, son las razones de su amplio uso en el genotipado animal.

#### *1.1.2.2 Características de los SNPs como marcadores moleculares.*

Los SNPs son cambios en una única base en una posición determinada en la secuencia de DNA, generalmente con dos posibles nucleótidos (2 alelos) en esta posición. Para que esta alternativa en aquella posición pueda ser caracterizada como "SNP válido" (i.e. como marcador molecular para genotipado), el alelo con menor frecuencia debe tener una representación de a lo menos un 1% a lo largo de la población. De no ser así se le denomina variante de sitio único (Single Nucleotide Variant, SNV).

En la práctica los SNPs son usualmente bi-alélicos, dada la baja frecuencia de ocurrencia de sustituciones de un nucleótido que originan los SNPs, estimadas entre $1 \times 10^{-9}$ y $5 \times 10^{-9}$por nucleótido por año en alguna posición neutral para mamíferos (Li et al., 1981; Martínez-Arias et al., 2001). Es por esto que la probabilidad de ocurrencia de dos sustituciones independientes en una misma posición es muy baja ($\sim 10^{-36}$, considerando un genoma del orden de $10^{9}$). Otro

motivo es la tendencia sesgada de ocurrencia de mutaciones, generando prevalencia de dos tipos de sustituciones.

De acuerdo a su estructura molecular, las bases nitrogenadas pueden ser agrupadas en dos categorías las pirimidinas (derivadas de la pirimidina C4H4N2:citosina, timina y su equivalente en RNA uracilo) y las purinas (derivadas de la purina C5H4N4:adenina y guanina).Tomando esto en consideración, podemos clasificar las sustituciones en dos grupos: transiciones (ti), la sustitución de purina por purina (A↔G) o pirimidina por pirimidina (C↔T); y transversiones (tv), la sustitución de una purina por pirimidina o viceversa (A↔C, A↔T, G↔C, G↔T).

En un sistema donde las mutaciones puntuales ocurren de manera aleatoria el radio de transiciones sobre transversiones tiende a 0,5, dado que tenemos la mitad de posibles transiciones (A↔G y C↔T = 4 posibilidades) que de transversiones (A↔C, A↔T, G↔C, G↔T = 8 posibilidades) (Strachan and Read, 2004).En la práctica se observa una notoria tendencia a una mayor ocurrencia de transiciones que transversiones, generando una prevalencia de transiciones a lo largo del genoma. Las implicancias prácticas de este sesgo en la proporción ti:tv se retomarán en mayor detalle en 1.3.3.1.3.

El principal beneficio de los SNPs como marcadores moleculares deriva de la necesidad de paneles de marcadores genéticos de alta densidad para los estudios de rasgos multifactoriales, y el reciente progreso en su detección y las técnicas de genotipado animal como laselección por desequilibrio de ligamiento.

### *1.1.3 Desequilibrio de ligamiento y estudios de asociación.*

Si existen marcadores moleculares de DNA disponibles estos pueden ser utilizados para determinar si la variación a nivel molecular (variación alélica en el locus del marcador y su mapa de ligamiento), se encuentra ligada a la variación del rasgo

cuantitativo. De ser así, el marcador se encuentra ligado o en el mismo cromosoma que el QTL el cual presenta variantes alélicas causantes de la variación delrasgo. Si se dispone de un número reducido de marcadores por genoma, entonces las asociaciones fenotípicas de estos marcadores se mantendrá sólo dentro de cada familia y por un número limitado de generaciones, debido a las recombinaciones históricas posibles entre los marcadores y los QTLs(Hayes, 2007).

Otra dificultad de este acercamiento es que a menos que se cuente con un gran número de progenie por familia, los QTLs mapearán con intervalos de confianza (CI) amplios en el genoma (Figura 2), resultando en una asociación pobre de corta perduración al través de las generaciones.

$$CI_{95\,\%} = \frac{G}{kN\delta^2}$$

**Figura 2. Ecuación para estimación del Intervalo de Confianza de ubicación de QTL(Darvasi and Soller, 1997).**
*N representa el número de individuos genotipados, δel efecto de sustitución del alelo (i.e. el efecto de obtener una copia extra del QTL) en unidades de desviación estándar residual, k el número de parentales informativos por individuo, y G que es el tamaño del genoma en centi-Morgan (cM).*

No obstante, si se dispone de un conjunto denso de marcadores, una alternativa a lo anteriores la utilización del principio de desequilibrio de ligamiento (Linkage Disequilibrium LD, asociación no aleatoria de alelos entre dos loci), explotando las asociaciones entre marcadores y QTLs (haplotiposmarcados que contengan al QTL).

El desequilibrio de ligamiento puede ser entendido como una medida de la desviación de las frecuencias de ocurrencia de los haplotipos en una población predichas por aleatoriedady puede ser medido como un inverso de la distancia estimada en "centimorgans"abreviado cM (donde 1cM corresponde a la distancia a la que la recombinación esperada entre dos loci en una generación es 0,01) e.g. si

el locus A tiene 2 posibles alelos A1 y A2 y el locus B a su vez tiene B1 y B2, tenemos 4 combinaciones de alelos (haplotipos) posibles en la población: A1_B1, A1_B2, A2_B1 y A2_B2. Si cada alelo tiene igual probabilidad de ocurrencia, por aleatoriedad esperaríamos encontrarlos representados un 0,25 c / u en la población.

El mapeo por LD requiere que un marcador se encuentre en LD con un QTL en una población y que este ligamiento persista a lo largo de las generaciones, paralo que deberán encontrarse a una distancia "corta" en el genoma (~ 0 cM) i.e.fuertemente ligados.

Una de las técnicas que utiliza LD para el mapeo de QTLs es el estudio de asociación de genoma completo (GWAS),que consiste en la examinación de un pool de marcadores, específicamente polimorfismos de un nucleótido (SNPs), y su LD con el QTL asociadoalos rasgos determinados (i.e. enfermedad, rasgos fenotípicos, etc.). Este análisis comparativo se realiza por ensayo de los genomas contra un chip o arreglo que contiene un conjunto de marcadores SNPscapaces de marcar efectivamente uno o más haplotipospresentes en el genoma, denominados SNPs reales o True-SNPs en inglés (Wang et al., 1998). Comparando grupos de individuos que presentan el rasgo cuantitativo en diferentes magnitudes -- idealmente un grupo de individuos que presenta el rasgo y un grupo control que no presentan el rasgo --,se puede determinarun subconjunto de SNPscon alto grado de asociación estadística al rasgo evaluado.Éstos pueden ser valorizadospara seleccionaren basealsubconjunto de genes o loci en LD con estos marcadores, que están asociados más fuertementesobre el rasgo en cuestión (QTLs).

El uso de un conjunto de marcadores de alta densidad aumenta la precisión en la determinación de QTLs además de permitir la caracterización de la heredabilidad e implicancias genéticas de rasgos que no son trazables mediante otras técnicas.

GWAS ha tenido un gran éxito en el mejoramiento de múltiples especies de importancia económica. Algunos ejemplos de esto son: en cerdo se generó un panel de 60 K marcadores moleculares de polimorfismos de una base (SNPs) que permite la correcta identificación de parentales(Rohrer et al., 2007);en vacuno por lo menos 2 compañías de crianza han lanzado al mercado familias de toros para su uso comercial (Hayes et al., 2009), basadas en criterios de selección por marcadores; y en aviares se identificaron QTLs involucrados en el síndrome de hipertensión pulmonar (ascitis), utilizando microsatélites como marcadores(Closter, Elferink, et al., 2010). Este éxito es atribuible tanto a las mejoras en las técnicas de biología molecular así como al desarrollo de herramientas bioinformáticas capaces de manejar y analizar la gran cantidad de datos generados de forma eficiente.

El objetivo principal de la examinación por GWAS es la determinación de QTLs. El análisis de QTLs generalmente procede en tres etapas: 1) Determinar las contribuciones relativas de los factores genéticos vs ambientales en las variaciones fenotípicas de una población, 2) Identificar las ubicaciones cromosómicas de estos loci y 3) Determinar los genes asociados a él o los rasgos y como variantes de estos genes influyen sobre el fenotipo (Lynch et al., 1998). La precisión y poder estadístico con el que estas problemáticas biológicas pueden ser resueltas ha mejorado notablemente en los últimos años, gracias al avance en las tecnologías de genética molecular las que han permitido determinar un mayor número de polimorfismos génicos y mutaciones, permitiendo la creación de mapas de ligamiento génico y búsqueda de QTLs con mayor eficiencia(Mardis and others, 2008; Reis-Filho and others, 2009; Schuster, 2008). Los chips para genotipado

comerciales actualmente son capaces de ensayar hasta un millón de polimorfismos (SNPs) a lo largo del genoma (Chicurel, 2002).

Algunos de los modelos estadísticos más utilizados para la determinación de la asociación marcador-QTL son: Test de Chi-cuadrado (utilizado para la evaluación puntual de un locus y las implicancias de sus alelos), regresión logística (asocia rasgo con múltiples factores genéticos y estos pueden tener efectos cruzados sobre el rasgo), Test de Fisher (para muestras de poblaciones pequeñas), Test de Tendencia de Cochram-Armitage (utilizado para los casos de alelos codominantes i.e. de expresión conjunta en caso de individuos heterocigotos).

### *1.1.4 Selección asistida por marcadores y selección genómica.*

Luego de la identificación de los marcadores específicos se realiza una selección de los animales asistida por marcadores (Marker Assisted Selection, MAS), esta puede ser realizada mediante el uso de marcadores en equilibrio de ligamiento con el QTL (LE-MAS, selección negativa), marcadores en desequilibrio con el QTL (LD-MAS, selección positiva), o bien utilizando la mutación causante del efecto QTL (Gene-MAS). Las tres técnicas son actualmente utilizadas en la industria ganadera para la selección de distintos rasgos de interés con resultados positivos(Bennewitz et al., 2003; Boichard et al., 2002; Dekkers, 2004; Plastow et al., 2003).

Sin embargo, resultados obtenidos en GWAS en ganado y humanos han llevado a concluir que la selección mediante un número limitado de QTLs(LE-MAS, LD-MAS, Gene-MAS), generalmente sólo captura una porción de la varianza genética total de rasgos complejos como productividad, probablemente dado que no considera la gran mayoría de genes involucrados en la determinación delos QTLs con efectos pequeños sobre el carácter (verFigura 1), y esto sumado al importante efecto de las secuencias no codificantes sobre la regulación del carácter(Djebali et al., 2012;

Neph et al., 2012; Thurman et al., 2012).Esto ha llevado al desarrollo de técnicascapaces de rastrearla gran mayoría de QTLs presentes en el genoma. Esto puede ser realizado mediante el método llamado selección genómica(Meuwissen et al., 2001a), que consiste en la división del genoma en segmentos cromosómicos y el posterior rastreo de cada segmento. La selección genómica utiliza un panel de marcadores de alta densidad de genoma completo, generando una alta probabilidad (~99,9%) de que todos los QTLs presentes en el genoma estén en desequilibrio de ligamiento con al menos un marcador presente en el panel (*SNPchip*). La selección se basa en los llamados valores de pedigríestimados o GEBV por sus siglas en inglés (GEBV, Genomic Estimator of Breeding Value)(Sargolzaei et al., 2009)que son predichos como la suma del efecto de estos SNPs o haplotipos a lo largo del genoma, más un factor ambiental aleatorio ponderado sobre los QTLs(Hayes et al., 2009).

Los estudios de asociación genómica constituyen una herramienta útil para el desarrollo de nuevas tecnologías y para el análisis y resolución de problemáticas biológicas. Para realizar una selección genómica eficiente es esencial contar con una base de datos que contenga el máximo de marcadores con asociaciones conservadas con variantes genómicas heredables y de significancia estadística (SNPs reales).

#### *1.1.4.1 Descubrimiento de SNPs para selección genómica.*

Al momento de elegir marcadores para selección genómica debemos considerar que estos alelos y sus asociaciones fenotípicas se conserven al través de las generaciones. El principio de Hardy-Weinberg (Stern, 1943) postula que ambos alelos y sus frecuencias genotípicas en una población permanecen constantes de generación en generación a menos que influencias perturbadoras específicas sean introducidas. Por lo que el marcador ideal con frecuencias alélicas estáticas asume:

alelos sin mutaciones, ni migraciones o migración nula (no hay intercambio de alelos entre poblaciones), tamaño poblacional infinito y presión selectiva nula a favor o en contra de cualquiera de los genotipos. En una población real siempre existirán influencias perturbadoras que desvíen a nuestros alelos del equilibrio de Hardy-Weinberg, no obstante el nivel de conservación y la representación en la población de cada alelo pueden ser indicadores de utilidad al momento de dirimir entre marcadores (ver Tabla IVy Tabla V).

Existen múltiples acercamientos para la búsqueda de SNPs, algunos actualmente utilizados para el genotipado.Los métodos principales se basan en la comparación de secuencias locus-específicas, generadas a partir de diferentes individuos. El método más simple, cuando se tiene como objetivo una región que contiene genes candidatos, consiste en el secuenciamiento directo de productos de PCR obtenidos a partir de diferentes individuos. Sin embargo, este acercamiento tiende a tener un elevado costo a gran escala ya que se requiere de primers locus-específicos para cada una de las regiones a evaluar. Adicionalmente si se secuencian individuos heterocigotos para el SNP en cuestión, éste se observará como un doble pick que no será distinguible de un error de secuenciamiento (ver Figura 3).

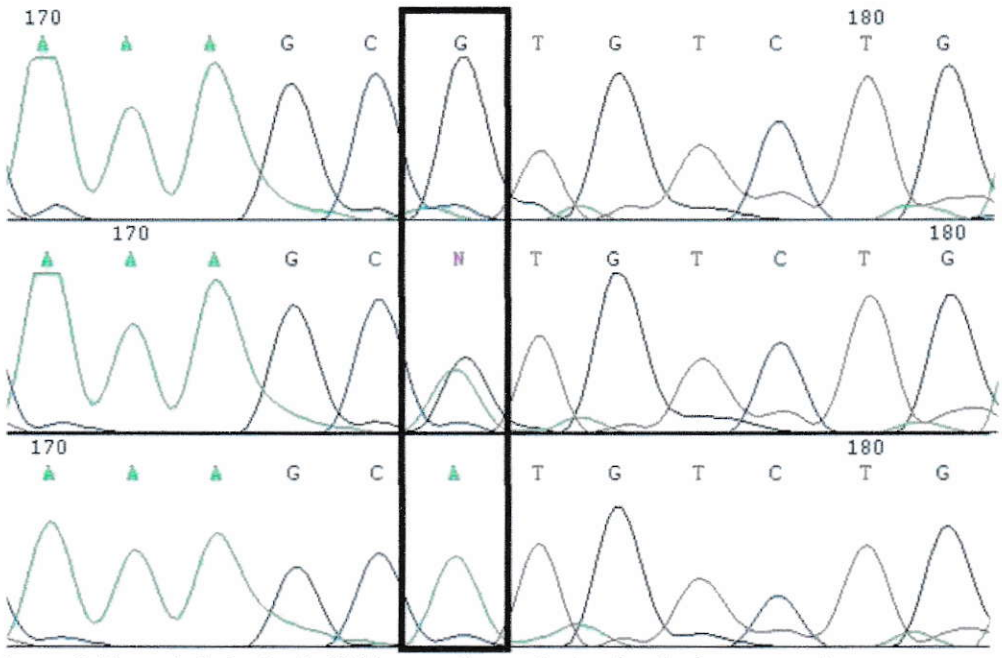



**Figura 3. Descubrimiento de SNPs por alineamiento simple de secuencias.**
*En el recuadro se destaca un polimorfismo donde el segundo individuo es heterogéneo para el SNP siendo imposible diferenciarlo de un artefacto de secuenciamiento. La primera muestra representa un individuo AA, la segunda muestra un individuo AG y la última un individuo GG.*

Estos problemas se pueden solucionar mediante la utilización de individuos que presenten alta homocigocidad o secuenciamiento de bibliotecas de clones.

Una técnica ampliamente utilizada es la obtención de SNPs a partir de la comparación de secuencias de fragmentos de DNA expresado (o ESTs por su sigla en inglés, Expresed Sequence Tag), provenientes de múltiples individuos.Aunque estas técnicas han sido exitosas en la búsqueda de SNPs (Kim et al., 2003), se limitan exclusivamente a una porción de las regiones codificantes del genoma; mientras que,como se explicó en el párrafo anterior,para una selección eficiente se requiere incluir marcadores distribuidos con una densidad homogénea a lo largo de todo el genoma.

Una alternativa a la anterior es la comparación de fragmentos de secuencias genómicas de múltiples individuos. Estas técnicas ha demostrado ser efectivas en la búsqueda de marcadores eficientes para la selección genómica para múltiples

organismosde importancia económica (Hyten et al., 2010a, 2010b; Sánchez et al., 2009; Van Tassell et al., 2008; Wiedmann et al., 2008).Sin embargolas particularidades de cada genoma generan dificultades para la aplicación de estas y otras técnicas en existencia, específicamente las especies que presentan poliploidía o pseudo-diploidía como plantas y peces de cultivo su selección asistida eficiente mediante marcadores han representado un gran desafío (Dominik et al., 2010). Un caso emblemático para la industria chilena donde estas técnicas presentan dificultades son los salmónidos. El mejoramiento de estas especies es vital para el futuro de la industria acuícola, es por esto que en un esfuerzo internacional y a través de múltiples financiamientos, tanto privados como públicos se planea no solo obtener un ensamble completo de referencia del genoma de la especie *Salmo salar,*si no también caracterizar marcadores eficientes para su mejoramiento.

## *1.2 Búsqueda de marcadoresgenéticos en Salmónidos.*

### *1.2.1 Importancia económica de los salmónidos.*

Los salmónidos (salmón y trucha) constituyen un género de gran importancia tanto económica como científica. Se estima que la producción mundial sobrepasa las 1,4Mt anuales (© FAO - Fisheries and Aquaculture Information and Statistics Service, 2012), siendo Noruega y Chile los mayores productores a nivel mundial, con aproximadamente 1 M y 600 K toneladas de salmón del Atlántico (*Salmo salar*), respectivamente (Nystøyl, 2011). El cultivo de salmónidos en Chile constituye el mayor segmento productivo dentro de la acuicultura, siendo el salmón del Atlántico (*Salmo salar*) junto con el salmón coho (*Oncorhynchus kisutch*) las especies acuícolas de mayor producción. Por esto, un mejoramiento en la producción de estas especies es esencial para el crecimiento productivo a nivel país. Sin embargo la identificación de marcadores eficientes para el mejoramiento, ha

encontradomúltiples dificultades en el caso de los salmónidos, debido al alto nivel de duplicación observado en estas especies.

### *1.2.2 Duplicaciones genómicas en salmónidos y sus consecuencias.*

El genoma de los vertebrados mandibulados ha sufrido dos rondas (2R) de duplicación de genoma completo (WGD)(Ohno and others, 1970), que tomaron lugar luego de la emergencia de los urocordados y antes de la radiación de los mandibulados (Kasahara, 2007).Subsecuentemente los teleósteos sufrieron una tercera ronda adicional de WGD (3R-WGD) alrededor de 320-400 Millones de años atrás(Sato and Nishida, 2010) (Figura 4) y por último una cuarta ronda (4R-WGD), sufrida por la subfamilia salmonidae aproximadamente 20 – 120 Millones de años atrás(Moghadam et al., 2011) (ver Figura 5), por lo que actualmente encontramos al genoma de estas especies sumido en un proceso de re-diploidizaciónmediante perdida de segmentos del genoma, silenciamiento génico o por divergencia donde cada gen duplicado presenta diferentes patrones de expresión (Bailey et al., 1978).

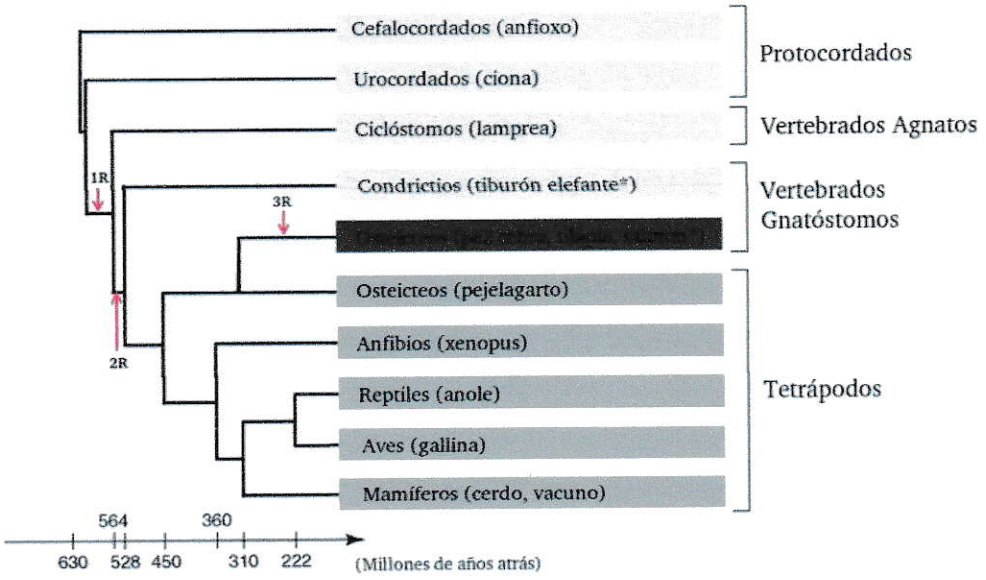



**Figura 4. Duplicaciones genómicas ocurridas a lo largo de la evolución de los salmónidos (Kasahara, 2007).**
*Representación de las 3 rondas de duplicación sufridas por los vertebrados a lo largo de su evolución. La primera (1R), alrededor de 564 Millones de años atrás durante la separación*

*de agnatos y mandibulados; Una segunda duplicación (2R), alrededor de 528 Millones de años atrás durante el surgimiento de los osteíctios; y una tercera (3R), alrededor de 310 Millones de años atrás, específica de los teleósteos. Entre paréntesis se incluyen ejemplos de especies que han sido secuenciadas o están siendo secuenciadas (*) para cada subfamilia.*

Aunque los mecanismos moleculares a través de los cuales estas duplicaciones ocurren permanece como un tema de amplio debate (Dehal and Boore, 2005), sus consecuencias pueden ser apreciadas en teleósteos como: secuencias parálogas con una identidad superior al 90 % (Moghadam et al., 2011); aproximadamente un tercio de genes con duplicados funcionales (Allendorf and Utter, 1973; Ohno and others, 1970); duplicación de marcadores (Allendorf, 1984); formación de complejos tetravalentes durante la meiosis masculina (Svärdson, 1945); el ligamiento aparente de loci no ligados debido a disociación no aleatoria de los complejos tetravalentes (pseudoligamiento) (May, 1980; Wright et al., 1980); patrones de segregación inusuales parcialmente tetrasómicos (Allendorf, 1984; Allendorf and Danzmann, 1997; Johnson et al., 1987; Jr et al., 1983); un número aumentado de brazos cromosómicos en comparación con otros peces de aleta (Phillips and Ráb, 2001); además de una elevada ocurrencia de elementos repetitivos (RE) y transposones (de Boer et al., 2007; Kido et al., 1994). Se estima que cerca del 50 % de las regiones codificantes del genoma de *Salmo salar* se encuentran en regiones duplicadas (Davey et al., 2001b; Utter et al., 1973).

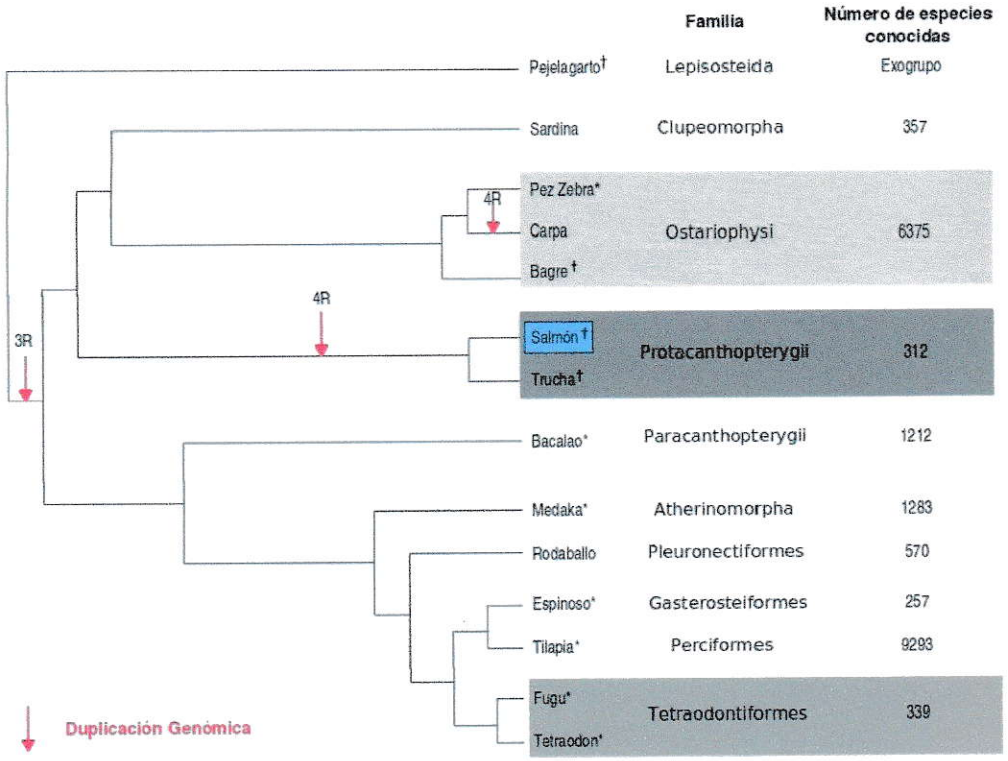



**Figura 5. Relaciones filogenéticas(Steinke et al., 2006) de especies de peces indicando los 7 genomas secuenciados de dominio público (*) (Flicek et al., 2010; Hubbard et al., 2002) (fugu, Takifugu rubripes; tetraodon, Tetraodon nigroviridis; medaka, Oryzias latipes; stickleback, Gasterosteus aculeatus; y zebrafish, Danio rerio), además de las especies que están actualmente siendo secuenciadas (†)(Davidson et al., 2010).**
*Las flechas rojas indican donde se habrían producido las rondas de WGD comenzando con la 3R. El Gar, un grupo de teleósteos ancestral se muestra en la base del árbol, a modo de grupo externo.*

Todo lo anterior genera problemas en la búsqueda de SNPs para selección genómica, al dificultarla diferenciaciónde SNPs realesy:1) artefactosde ensamblegeneradospor variaciones entre sitios parálogos del genoma que poseen una alta similitud, las denominadas variantes de secuencias parálogas o PSVs por su sigla en inglés (PSV, Paralogue Sequence Variants);2) polimorfismosreales presentes ensitios duplicadosdel genoma, denominados MSV por sus siglas en inglés (MSV, Multiple Sequence Variants).

Un PSV se forma cuando secuencias parálogas difieren en una o múltiples bases pero no hay segregación de ninguna de las dos secuencias (Figura 6). Otra fuente

de variación en genomas poliploides son las variantes de sitios múltiples (MSV) que a diferencia de los PSVs, segregan por la sustitución de una base en uno o ambos loci parálogos (Fredman et al., 2004) (Figura 6).

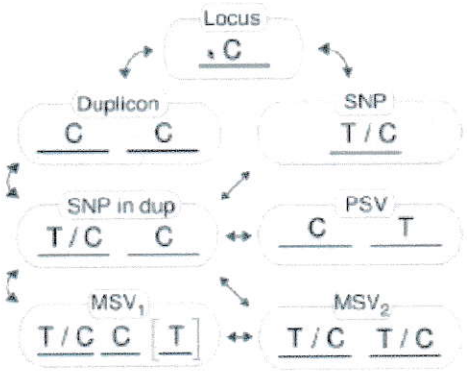



**Figura 6. Formación de polimorfismos (PSVs y MSVs) a partir de duplicaciones cromosómicas(Fredman et al., 2004).**
*Secuencia de cambios evolutivos desde una base monomórfica a un MSV polimórfico. Las flechas indican eventos tales como mutaciones, fijación, duplicación, deleción y conversión génica. La mayoría de estos procesos son reversibles.*

Por su parte los MSVs pueden ser subcategorizados como: MSV-3s cuando solo uno de los duplicones o secuencias parálogas es variante;y MSV-5s cuando ambas secuencias parálogas presentan variantes independientes(Lien et al., 2011).

Podemos resumir las dificultadesque los puntos anteriores representan para la selección genómica en salmones como:

1)Los PSVs no constituirían marcadores reales ya que al no ser segregantes no permitirían diferenciar entre individuos, resultando en conjuntos de marcadores que presentan regiones con alta densidad de SNPs de los cuales un alto porcentaje representa artefactos como los ya nombrados y por lo tanto, son ineficientes para el marcaje de QTLs y selección genómica.

2) Los MSVs a pesar de ser segregantes, la imposibilidad de diferenciarlos de PSVs *a priori*para su caracterización, y la dificultad para su selección en el caso de

presentar 2 variantes independientes (MSV-5s), dificulta su uso en la selección genómica aumentando el valor y tiempo asociado alaobtención de conjuntos de marcadores segregantes de alta densidad, efectivos para el genotipado(Lien et al., 2011). A pesar de esto, nuevos algoritmos matemáticos han sido desarrollados para la asignación de MSV-3s a la secuencia paráloga correcta y el descarte de MSV-5s, sin embargo todos estos acercamientos son *a posteriori*(i.e. luego del genotipado)(Gidskehaug et al., 2011), teniendo un costo y tiempo elevados.

3) El bajo nivel de conservación de las regiones del genoma con parálogos funcionales con alta identidad, conlleva a la obtención de marcadores que se desvían del HWE y por lo tanto no son de utilidad para la selección asistida. Por otra parte la exclusión unilateral de este tipo de regiones para la búsqueda de marcadores genera amplias regiones sin marcadores eficientes, imposibilitando la aplicación de técnicas como la selección genómica.

Estos efectos se han traducido en dificultades para la aplicación de metodologías tradicionales de búsqueda y evaluación de SNPs aplicadas a la selección genómica, así como para la generación de una secuencia de referencia para el genómade salmónidos. Una evaluación empírica del último SNPchip desarrollado para esta especie demostró que menos de un 20% de la totalidad de SNPs presentes en el conjunto representarían regiones únicas del genoma, de estos 350 se alejaban de la distribución de Hardy–Weinberg(Stern, 1943)($p<0,05$) (i.e. estaban presentes en zonas poco conservadas, de alta incidencia de mutaciones) y 9004 pares tenían valores deLD mayores a 0,175, mientras que aleatoriamente se esperaban alrededor de 2000, lo que nos demuestra que una gran cantidad de SNPs se heredarían de manera conjunta (Dominik et al., 2010); mostrando la necesidad de la determinación de un nuevo conjunto que excluya PSVs, además de elementos repetitivos y transposones presentes en el genoma del salmón.

### *1.2.3 Avances en la caracterización del genoma del salmón del Atlántico.*

Dadas las dificultades para el secuenciamiento del salmón y su gran interés tanto económico como científico es que se formó una colaboración internacional (de la que este proyecto es un usuario), para el secuenciamiento del genoma del salmón (International Collaboration to Sequence the Athlantic Salmon Genome, ICSASG), compuesta inicialmente por investigadores, empresas y organismos de financiamiento gubernamentales y privados de Chile, Noruega y Canadá (incluyendo: Agencia de desarrollo económico de Chile, InnovaChile, Consulado de Investigación de Noruega, Fondo de Investigación de Pesca e Industria Acuícola de Noruega). A través de éste y otros esfuerzos anteriores se han logrado enormes avances en la caracterización del genoma del salmón del Atlántico.Se determinó el valor constante (denominado C-value, cantidad de DNA contenida en el núcleo del gameto expresada en pico-gramos), para el salmón del Atlántico estimado alrededor de 3,27 pg (Hardie and Hebert, 2003), lo que se traduce en un tamaño de total de genoma de ~3.2Gb. El contenido de G+C del salmón del Atlántico (44,4%)(Bucciarelli et al., 2002). Se determinó que a pesar de ser un genoma de tamaño y composición similar a vertebrados de sangre caliente, se encuentra desprovisto de estructuras isocóricas (regiones de DNA mayores a 300 Kb, con un gran grado de uniformidad en el contenido de GC), como es de esperar en genomas de peces de agua fría (Bernardi, 1989). Se ha sugerido que el ancestro diploide de los salmónidos poseía un cariotipo con 48 cromosomas acrocéntricos, resultando en 96 cromosomas acrocéntricos luego de la duplicación genómica (Phillips and Ráb, 2001). Comparaciones entre los cariotipos de múltiples especies de salmónidos han reveladoque una enorme cantidad de reordenamientos cromosómicos (fusiones e inversiones) han venido ocurriendo e lo largo de los distintos linajes de salmónidos desde la ocurrencia de la última WGD (Phillips and Ráb, 2001). Distinto número de cromosomas entre las familias europeas y las

americanas (29 en *Salmo salar*de Europa y 27 en los americanos) (Lubieniecki et al., 2010).

Todo lo anterior gracias a la obtención de un individuo doble haploide producido mediante androgénesis mitótica(Seguí-Simarro and Nuez, 2008) de sexo femenino apodado "Sally". Su naturaleza homocigótica fue verificada mediante búsqueda de polimorfismos en loci de ~70 microsatélites. El cariotipado de Sally develó 29 cromosomas haploides, lo que se corresponde con lo esperado para el salmón del Atlántico noruego (Roberts, 1970). No se observaron reordenamientos cromosómicos aparentes; sin embargo, Sally aparenta ser un mosaico con un ~30% de células haploides y el resto diploides (Davidson et al., 2010). Se eligió un individuo de sexo femenino por ser éste el sexo homogamético (Davidson et al., 2009).Todo lo anterior apunta a facilitarel secuenciamiento y ensamble de una secuencia referencia para esta especie.

Adicionalmente en estos momentosse dispone de más de 200 librerías públicas de cDNA para distintos tejidos y estados de desarrollo del salmón del Atlántico (Adzhubei et al., 2007; Andreassen et al., 2009; Davey et al., 2001b; Hagen-Larsen et al., 2005; Koop et al., 2008; Leong et al., 2010; Martin et al., 2002; Rise et al., 2004; Tsoi et al., 2004). Hasta el 26 de Abril del 2012 existian495.211 ESTs públicos para *Salmo salar* en ESTdb de NCBI (Boguski et al., 1993). Se construyó una librería pública de cromosoma artificial bacteriano (o BAC, Bacterial Artificial Chromosome), de salmón del Atlántico denominada CHORI-214, a partir de un individuo macho proveniente de una cepa de cultivo de Noruega(Thorsen et al., 2005). Mediante la huella genética de la digestión por HindIII de CHORI-214 se creó el primer mapa físico del genoma, constituido por 223.781 BACs en ~4.565 contigs y 33.217 singletons (Ng et al., 2005).También se cuenta con una base de datos de elementos repetitivos específicos de salmónidosdisponible públicamente y

mantenida por el consorcio de investigación genómica en salmónidos o cGRASP (cGRASP, consortium Genomics Research on All Salmonids Project)(Secko et al., 2007). Además en el año 2011 el laboratorio chileno Math$^{omics}$(www.mathomics.cl) generó una base de datos pública llamada SalmonDB que contiene una completa caracterización de la información genómica anterior e incluye Unigenes (transcriptoma) generados a partir de estos ESTs públicos, candidatos SNPs, RE, entre otros (Di Génova et al., 2011).

A pesar de la acumulación de secuencias, hasta la fecha no existe un genoma de referencia final, esto dada la naturaleza repetitiva del genoma del salmón del Atlántico y las dimensiones de los elementos repetitivos de su genoma (~1.500 pb) (de Boer et al., 2007).Lo que hace necesario la utilización de técnicas de secuenciamiento y ensamble específicas. Sin embargo múltiples laboratorios del ICSASG comparten ensambles preliminares con altas representaciones generados a partir de las secuencias públicas provenientes de Sally(Ng et al., 2005)(ver1.3.1), de los cuales AGDK01 fue publicado el 12 de Octubre del 2011 en NCBI(Davidson et al., 2010).

Por otra parte, en los últimos 3 años se ha logrado caracterizar múltiples conjuntos de SNPs reales en salmón mediante la utilización de distintas técnicas.Se lograron caracterizar 112 SNPs reales en salmón del Atlántico a partir decDNA secuenciados a tamaño completo (FLIc, Full Length Sequenced cDNA) comparando secuencias codificantes contra no codificantes y utilizando como indicador su nivel de conservación (Andreassen et al., 2010).En el año 2011 se generó un mapa de ligamiento extenso mediante genotipado de 3.297 individuos para cultivo noruegos provenientes de 143 familias, utilizando 5.768 SNPs (Lien et al., 2011), donde 2.696 de estos marcadores fueron obtenidos a partir de ESTs o genes asociados y los marcadores remanentes se obtuvieron por secuenciamiento aleatorio de

fragmentos genómicos, recolectados mediante la técnica de librerías de representación reducida (RRL, Reduced Representation Libraries). Elarreglo del SNPchip de ~6 K se generó en Infinum® de Illumina®; los resultados del genotipado fueron analizados mediante el paquete beadarrayMSV de R (Gidskehaug et al., 2011), el cual permite diferenciar MSV-3s, MSV-5s y True-SNPs de acuerdo a sus patrones de genotipado (ver Figura 7). Se identificaron un 21% de los marcadores como MSVs mapeando de manera heterogénea a lo largo del genoma, la homología de estas regiones fue chequeada por mapeo contra una referencia preliminar(Gidskehaug et al., 2011),y develóclaramente patrones de regiones duplicadas a lo largo del genoma.Además se observaron diferencias en los patrones de recombinación entre sexos, con machos mostrando una menor taza de recombinación (Lien et al., 2011), por lo que los marcadores fueron genotipados en su mayoría en hembras.

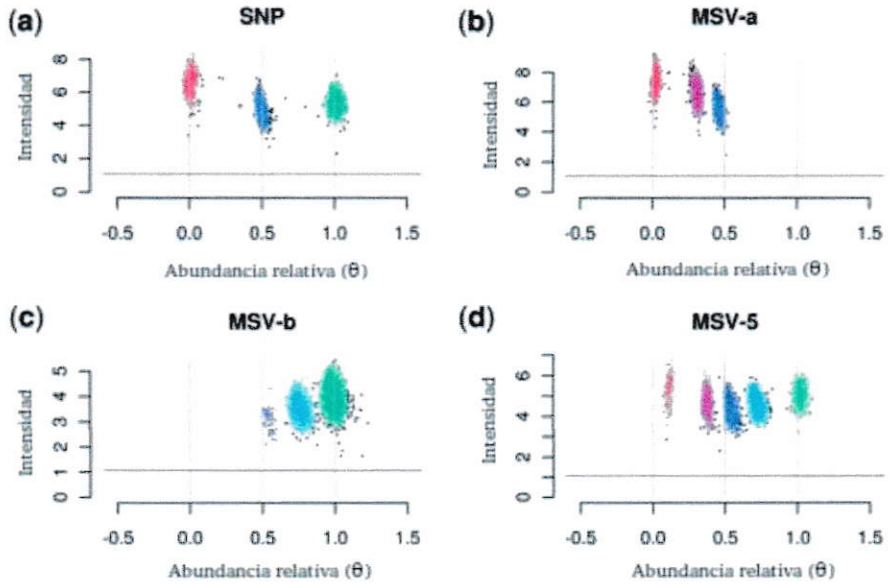



**Figura 7. Ejemplos de patrones de agrupamiento observados durante el genotipado según la naturaleza de los marcadores utilizados(Gidskehaug et al., 2011).**
Los colores rojo y verde representan a los dos alelos posibles (e.g. A y B)el color azul representa 50% alelo "rojo" + 50% alelo "verde"(e.g. AB), el color purpurarepresenta 25% alelo "rojo" + 75% alelo "verde" (e.g. AA AB) y el color azul marino representa 75% alelo "rojo" + 25% alelo "verde" (e.g. AB BB) (a) *Un SNP de una región diploide.Rojo representa AA, azul representa AB y verde representa BB. (b) Un MSV-3s donde un locus es variable y*

*el otro es homocigoto para el alelo "rojo". (c) Un MSV-3s donde un locus es variable y el otro es homocigoto para el alelo "verde". (d) Un MSV-5s donde se observan variaci ón en ambos loci.*

A pesar de la disponibilidad de estos marcadores de forma pública y los múltiples esfuerzos para generar selección eficiente en salmones(Li et al., 2011; Norman et al., 2012),hasta el momento sólo se ha caracterizado efectivamente un QTL en salmón del Atlántico relacionado con ganancia de peso, utilizando un arreglo de 6,5 K SNPs (Gutiérrez et al., 2012)sobre un grupo de individuos de una población para cultivo de Canadá.

Aunque este arreglo fue exitoso para este grupo de individuos se requiere aún de un arreglo de mayor densidad (~60 K) para generar selección genómica eficiente, capaz de excluir marcadores no segregantes (PSVs) e idealmente categorizar los marcadores obtenidos según su nivel de conservación para evitar el uso de regiones de baja conservación dentro de lo posible en el arreglo final (i.e. evitando MSV-5s). Además cabe considerar que la generación de un arreglo de SNPs que incluya muestras de ejemplares chilenos, será más efectivopara la selección de éstos, al incluir marcadores propios de los haplotipos presentes en las cepas de la región.

## *1.3 Propuesta para la obtención de variantes en especies con múltiples duplicaciones genómicas evolutivamente recientes, utilizando como especie modelo "Salmo salar".*

La búsqueda de variantes genómicas se basa en la comparación de secuencias de múltiples individuos, y la identificación de diferencias entre estas; esto se puede llevar a cabo comparando múltiples secuencias codificantes cortas como ESTs (DNA codificante) o RRL (DNA genómico) de múltiples individuos o comparando contra un genoma de referencia de la especie. La primera aproximación, aunque efectiva para la búsqueda de marcadores moleculares (Andreassen et al., 2010),

debido a las particularidades del genoma del salmón, introduce múltiples variaciones no validas causadas por PSVs (Dominik et al., 2010).Una manera de eliminar los PSVs *a posteriori*es evaluando el conjunto de marcadores contaminado contra una población a genotipar y observar los patrones de agrupamiento de los marcadores (ver Figura 7) con lo que se pueden clasificar de manera visual (Lien et al., 2011).Sin embargo, esta aproximación requiere de una gran inversión por la generación de múltiples SNPchips a medida que se generan las caracterizaciones, haciendo necesaria una caracterización *a priori* de los marcadores a utilizar. Es por esto que se requiere de un genoma de referencia que nos permita descartar elementos repetitivos, parálogos y otros artefactos *a priori* para luego realizar identificación de variables.

Nuestra hipótesis es que se puede determinar el estado de conservación de ventanas móviles a lo largo del genoma, para así clasificar cada marcador según el estado de conservación del contexto genómico en el que se encuentre.

Tomando esto en consideración, para la determinación de este nuevo conjunto de marcadores para el genotipado eficiente de *Salmo salar*proponemos seguir un proceso similar al que se muestra en la Figura 8.

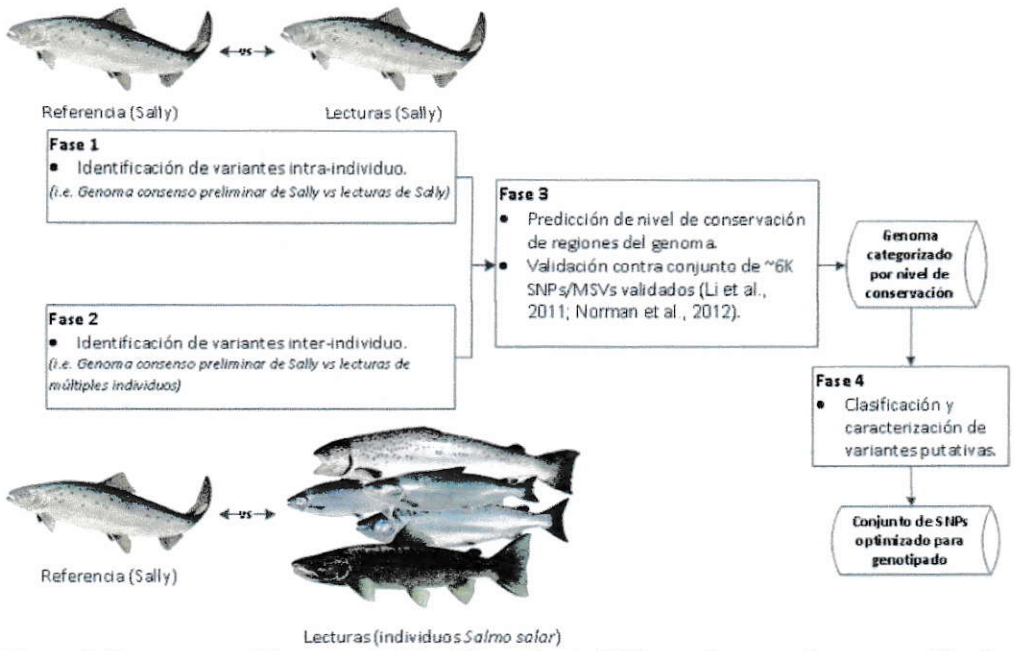



**Figura 8. Pasos sugeridos para la identificación de SNPs reales para la generación de un chip de genotipado eficiente (Elaboración propia).**
*El procedimiento global propuesto se compone de 4 fases mayores: Fase 1:Identificación de variantes intra-individuo (candidatos PSV/MSV). Fase 2: Identificación de variante inter-individuo (candidatos SNPs/DIPs/PSVs/MSVs). Fase 3: Determinación de indicadores de conservación de candidatos y clasificación deregionesdel genoma según sus niveles de conservación; Validación de la efectividad de las predicciones contra un conjunto de marcadores validados (Lien et al., 2011). Fase 4: Clasificación y caracterización de las variantes putativas, para el diseño de un conjunto de SNPs óptimo para el genotipado.*

Podemos resumir el protocolo propuesto en 4 fases principales:

1) La determinacióndevariantesentre las secuencias parálogas de un mismo individuo doble haploide(potenciales PSVs/MSVs) presentes en el ensamble preliminar representativo del genoma de referencia. Estas variantes serán denominadas variantes individuales (Individual Variants, IVs).
2) La obtenciónde variantes entre secuencias de múltiples individuos(potenciales SNPs/DIPs/MSVs/PSVs), las que denominaremos variantes de múltiples individuos (Multiple Individual Variants, MIVs).
3) La predicciónde nivel de conservaciónde distintas regiones del genoma por clusterización(ver1.3.3).

4) La categorización de las variantes putativas finales según el nivel de conservación de la región genómica en la que se encuentran cruzando los datos de IVs y MIVs obtenidas, produciendo así un conjunto de variantes mejoradas para el genotipado por selección genómica.

La eficiencia de esta aproximación puede ser evaluada contra un conjunto de variantes caracterizadas y validadas por genotipado (Lien et al., 2011). Todo lo anterior debe ser realizado luego de un filtrado de los elementos repetitivos y transposones presentes en el consenso y una limpieza de las lecturas crudas por enmascaramiento de estas secuencias; evitando variantes falsas.

Para facilitar la lectura en las secciones subsiguientes y evitar confusiones cabe resaltar que denotaremos a los candidatos PSVs/MSVs provenientes del mapéo de secuencias intra-individuo como candidatos "Individual Variants" o IVs, mientras que aquellos marcadores candidatos provenientes del mapéo de secuencias inter-individuos (que potencialmente pueden incluir posibles SNPs, DIPs, PSVs y MSVs) les llamaremos "Multi-Individual Variants" o MIVs.

El protocolo anterior tiene por objetivo la obtención de una librería de polimorfismos que incluya un conjunto de marcadores con un máximo de SNPs/DIPs reales de alta densidad, depurado y caracterizado obtenido utilizando cepas chilenas. Para la realización de este protocolo debemos considerar 3 metodologías: 1) la generación de un genoma de referencia preliminar; 2) la comparación de secuencias contra este genoma de referencia; 3) la identificación de regiones duplicadas en este genoma de referencia.

Este protocolo representa un acercamiento único para el filtrado *a priori* de conjuntos de marcadores para una selección genómica eficiente.

### *1.3.1 Generación de un genoma de referencia preliminar.*

En humano el secuenciamiento se realiza por mapeo de lecturas sobre un genoma de referencia público, lo que disminuye considerablemente los costos de secuenciación. Cuando no se dispone de secuenciamientos previos se debe convertir una colección de secuencias compuestas por pequeñas lecturas en ensambles consistentes, esto se denomina secuenciación *de novo* y se han escrito más de una docena de programas computacionales con este propósito.

El ensamble de un genoma consiste básicamente en la composición de las secuencias genómicas completas (o lo más completas posible), que componen a los cromosomas de un organismo, a partir de lecturas de fragmentos de DNA, esto requiere de algunas de las capacidades de computación y procesamiento más altas en el campo de la biología. La primera tecnología de secuenciamiento masiva fue el secuenciamiento Sanger ® (Sanger et al., 1977), capaz de producir fragmentos de DNA (lecturas) de hasta 1.000 pares de bases, las lecturas adyacentes generalmente se superponen por unos cuantos cientos de pares de bases. Esto para un genoma de 3,2 Gb significaría ensamblar un puzzle de a lo menos 64 millones de piezas (a una cobertura de 2X y tamaño promedio de lectura de 100 bp), de las cuales algunas piezas estarán ausentes, otras contendrán errores y algunas calzarán en más de un lugar en el genoma (repeticiones). Para compensar, los ensambladores requieren de un gran número de lecturas de una misma zona del genoma para facilitar el ensamble.

Un problema con la secuenciación Sanger es su elevado costo y tiempo de reacción, es por esto que se han generado nuevas tecnologías de secuenciamiento de lecturas cortas de alto rendimiento a un 0,1% del costo, denominadas "New Generation Sequencing" o NGS. Sin embargo estas tecnologías representan un

desafío mayor al generar lecturas de solo unos cientos pares de bases(Baker, 2012).

Para generar un ensamble se utilizan los datos de las lecturas secuenciadas y se organizan para encontrar superposiciones entre estas fusionándolas en una secuencia contigua denominada "contig". La longitud de estas secuencias contiguas estará limitada por secuencias repetitivas, polimorfismos, ausencia de ciertos fragmentos del genoma y errores de secuenciamiento. Para resolver estos problemas es que se han desarrollado tecnologías capaces de obtener lecturas pertenecientes a los dos extremos de un fragmento de DNA de largo conocido, que permita que ambos extremos se encuentren a una distancia significativa, esto se denomina "pair-end sequencing" (Golding et al., 2000). El largo del fragmento del centro (no secuenciado) viene determinado por la preparación de las librerías, variando entre cientos ("pair-end sequencing" (Golding et al., 2000)) hasta varios miles de pares de bases ("mate-pair sequencing"(Pevzner and Tang, 2001)). Si estaslecturas se ubican en dos contigs distintos, podemos deducir que estos son contiguos y podemos generar un "andamio" o "scaffold" entre ambos, deduciendo el tamaño del espacio que los separa (que denominaremos gap) (ver Figura 9).

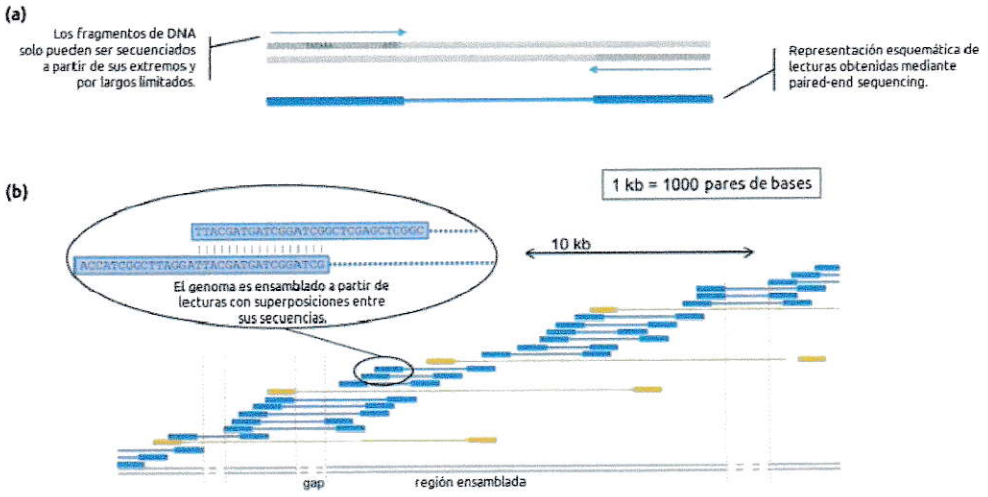



**Figura 9. Ensamble *de novo* utilizando paired-end sequencing.**
*(a) Muestra la técnica de secuenciamiento por paired-end (gris) y como sobrepasa las limitaciones en largo máximo de secuenciamiento en técnicas de NGS (azul). (b) Muestra la formación de un scaffold a partir de múltiples contigs de lecturas consecutivas (azul) y en gris la secuencia consenso resultante incluyendo sus gaps.*

Al no contar con un genoma de referencia de alta calidad, las secuenciaciones *de novo* pueden ser evaluadas en base a:el número de scaffolds y contigs requeridos para representar el genoma; la proporción de lecturas que fueron incluidas exitosamente en el ensamble;el largo absoluto de contigs y scaffolds en pb (sin incluir gaps); y la longitud de contigs y scaffolds relativa al tamaño estimado del genoma (representación). La medida más utilizada es N50, que consiste en el scaffold o contig más pequeño sobre el cual el 50% del ensamble se encuentra representado. Aunque esta medida por sí sola no siempre representa un indicador objetivo de la calidad del ensamble (Baker, 2012).

Dadas las dificultades que representa el ensamble de un genoma, se han generado esfuerzos mundiales como GAGE (Genome Assembly Gold-standard Evaluations)(Salzberg et al., 2012)y dnGASP (*de novo* Genome Assembly Assessment Project)(Earl et al., 2011), las que determinan indicadores estandarizados de la calidad de los ensambles para ser considerados válidos como genoma de referencia.

Un punto importante de considerar al momento de elegir un ensamblador es que el mejor algoritmo y sus parámetros óptimos vendrán determinados por las características del genoma del organismo a evaluar. Otro punto importante es que los mejores resultados se obtienen al incluir correcciones previas a las lecturas para eliminar errores de secuenciamiento (Salzberg et al., 2012).Para esto se requiere de la eliminación selectiva de secciones de acuerdo a su calidad de lectura. Este procedimiento se denomina trimmingy requiere contar con las calidades individuales de secuenciamiento de cada base que componen cada una de las lecturas secuenciadas.Éstas deben ser contenidas en un tamaño mínimo para facilitar su almacenamiento; el formato estándar que cumple con estos dos criterios es el llamado FASTQ (Cock et al., 2010),y consiste en una representación de la secuencia de amino ácidos de la lectura secuenciada como una secuencia de las letras A, C, T y G(basado en formato FASTA (Lipman et al., 1985)),seguido por una representación de la confianza con la que esa base fue asignada a la posición en cuestión. Esto se denomina calidad y existen dos escalas para calcularla, las cuales se muestran en la Figura 10.Adicionalmente existen múltiples estándares de representación de estos datos. Los estándares presentes en las lecturas utilizadas en este trabajo se resumen en la

Tabla I.

$$Q_{PHRED} = -10 \times \log_{10}(P_e)$$

$$Q_{Solexa} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)$$

**Figura 10. Ecuaciones para el cálculo de la calidad de secuencia.**
*Las ecuaciones representan los cálculos de calidad en base a la probabilidad de error tipo II según los dos estándares de calidad de secuencia.*

**Tabla I. Caracterización de los estándares de calidad de secuencia.**

| Descripción, denominación OBF | Caracteres ASCII | | Escala de calidad | |
|---|---|---|---|---|
| | Rango | Partida | Tipo | Rango |
| Estándar de Sanger `fastq-sanger` | 33-126 | 33 | PHRED | 0 a 93 |
| Solexa/Illumina<1.3 `fastq-solexa` | 59-126 | 64 | Solexa | -5 a 62 |
| Illumina 1.3+ `fastq-illumina` | 64-126 | 64 | PHRED | 0 a 62 |

Dadas las múltiples duplicaciones y los abundantes elementos repetitivos del genoma del salmón, se requiere del uso conjunto de librerías mate-pair y paired-end sequencing y secuencias Sanger,para la generación de un ensamble híbrido eficiente. El algoritmo ensamblador que ha mostrado los mejores resultados fue Celera (Levy et al., 2007; Miller et al., 2008; Myers, 2000) al que se le incorporaron múltiples correcciones por calidad especificamente desarrolladas para organismos con alto número de duplicaciones (expuestas en ICISB por Jason Miller).

Estos ensambles de referencia junto con lasbases de datos de elementos repetitivos específicos para salmónidos (Secko et al., 2007) e inespecíficos (Jurka et al., 2005), permiten el descarte de estos elementos repetitivos del genoma de *Salmo salar*, facilitando la identificación de variantes reales.

### *1.3.2 Contraste de secuencias contra genoma de referencia.*

Para el correcto contraste de secuencias de NGS contra un genoma de referencia es importante considerar incluir solo información real y relevante contenida en las secuencias crudas, i.e. debemos cuidar de eliminar: las secuencias de los vectores y/o adaptadores utilizados para la generación y secuenciamiento de la librería; y regiones de baja calidad que, por ende tienen probabilidades de estar incorrectamente secuenciadas. Al procedimiento de obtención de secuencias "limpias" a partir de secuencias "crudas" se le denomina "trimming" o recorte de secuencias (Li and Chou, 2004). Estos procedimientos generalmente se realizan

por separado con distintas herramientas o líneas de código personalizadas capaces de eliminar cada tipo de "impureza" por separado. Por esto el centro COMAV de la Universidad de Valencia desarrolló un pipeline automático basado en el lenguaje Python que junta estas herramientas estándar denominado clean_read (Blanca et al., 2011).

Luego de obtener la secuencia limpia, para poder comparar contra un genoma de referencia se debe encontrar la posición de esta secuencia en el genoma (mapeo) y determinar la posición de cada una de las bases de la secuencia con respecto a la referencia (alineamiento).

A lo largo de la historia de la bioinformática se han propuesto múltiples métodos de alineamiento de secuencias. Uno de estos métodos son las llamadas tablas de dispersión o "hashing" que fueron los primeros utilizados en alineadores de secuencias en formato FASTA, incluídos en el paquete FASTP (Pearson and Lipman, 1988), los cuales eran capaces de calcular alineamientos en menores tiempos que algoritmos basados en programación dinámica. En esta línea el paquete BLAST (Altschul et al., 1990), es el más ampliamente utilizado. Estas herramientas nos proveen de velocidad suficiente como para alinear ESTs producidos por secuenciadores basados en electroforesis capilar. Sin embargo las tecnologías de NGS, al proveernos de una cantidad de datos considerablemente mayor a gran velocidad y bajo costo, requieren de una velocidad de alineamiento mayor que no puede ser alcanzada por algoritmos como BLAST.

Para alcanzar estas demandas se han generado sobre 25 paquetes de programas diseñados para el mapeo y alineamiento de secuencias cortas de DNA a genomas de referencia. Sus algoritmos pueden ser clasificados en 2 categorías, los basados en tablas de dispersión (hashing) y los basados en compresión por ordenamiento

de bloques por sufijos, denominada Compresión de Burrows-Wheeler (BWT, Burrows-Wheeler Transform) (Burrows and Wheeler, 1994). La velocidad de este último grupo está determinada por el número de incongruencias entre las secuencias alineadas y el genoma de referencia.Tomando en consideración que las lecturas provenientes de secuencias expresadas solo contendrán fragmentos codificantes del genoma (secuencias exónicas), los alineadores basados BWT seránmenos eficientes en el alineamiento de estos. Es por esto que el tipo de algoritmo de alineamiento a utilizar será dependiente del origen de las lecturas de secuencias, usando algoritmos basados en BWT para secuencias genómicas y algoritmos basados en hashing para lecturas de expresión. Se han realizado múltiples esfuerzos para generar un algoritmo de alineamiento universal, sin embargo hasta el momento solo ha sido posible sacrificando la precisión del alineamiento (Hach et al., 2010).

Comparaciones entre las últimas versiones de alineadores de estas dos ramas disponibles han concluido que BWA(Li and Durbin, 2009, 2010) y GMAP (Wu and Watanabe, 2005) arrojarían los mejores resultados. BWA consta de 3 algoritmos independientes permitiendo el alineamiento y mapeo de: secuencias genómicas de 70 o más bp (NGS sequencing); secuencias más cortas (óligos de marcado); y secuencias con "gaps" frecuentes (i.e. datos transcriptómicos, mediante hashing ssaha2). Sin embargo GMAP nos presenta con la mejor alternativa para el mapeo de datos transcriptómicos produciendo el mínimo de falsos positivos, sacrificando un mínimo de velocidad(Wu and Watanabe, 2005).

Existen múltiples formatos de salida para estos alineamientos, con distintas representaciones dela información, generalmente cada paquete de alineamiento en existencia define su propio formato de salida. Es por esto que el proyecto de colaboración internacional 1.000 Genomes (Siva, 2008), generó el formato genérico

SAM (Sequence Alignment/Map) (Li et al., 2009) con 5 objetivos principales: 1) constituir un formato suficientemente flexible como para almacenar información producida por múltiples programas de alineamiento de secuencias, 2) ser suficientemente simple como para permitir la conversión entre distintos formatos pre-existentes, 3) permitir un tamaño de archivo compacto (el formato SAM puede ser comprimido en código binario, identificándolo con la extensión BAM, Binary Alignment Map), 4) tener bajos requerimientos de memoria volátil para la computación de operaciones comunes y 5) permitir la indexación de los alineamientos por posición genómica mejorando la eficiencia de visualización de los datos. Además incluye una serie de herramientas para su manipulación y análisis denominadas SAMTools(Li et al., 2009).

Estos archivos de alineamientos pueden ser analizados en búsqueda de variantes entre las secuencias las que se almacenan en el formato genérico VCF (Variant Call Format) (Danecek et al., 2011), también desarrollado por el proyecto 1.000 Genomes con similares características y herramientas para su manipulación y análisis (VCFTools(Danecek et al., 2011)).

La mayoría de los formatos de salida utilizados en la bioinformática son reportes en forma de texto plano, por lo que lenguajes de programación específicos para el análisis de este tipo de reportes, como Perl (Wall and others, 1994) y Python (Sanner, 1999), han sido ampliamente utilizados. Estos lenguajes tienen una serie de ventajas de las que la más importante para el desarrollo colectivo es su capacidad de programar de manera modular orientada a objetos, haciendo posible la reutilización de algoritmos, lo que ha generado grandes librerías específicas para el área de la biología, como Bioperl (Stajich et al., 2002) y Biopython (Chapman and Chang, 2000). La gran ventaja de Bioperl sobre Biopython fue su incorporación al

proyecto genoma humano (Stein, 1996), lo que permitió el rápido crecimiento de las librerías públicas disponibles, siendo predominante hasta el día de hoy.

La gran plasticidad del lenguaje Perl ha permitido su desarrollo como lenguaje de pegado en la biología, sin embargolas tecnologías de NGS exigen una inmensa capacidad de procesamiento para el análisis del gran número de secuencias y alineamientos, haciendo necesario un lenguaje que permita paralelizar y distribuir de manera eficiente los análisis entre múltiples procesadores. Con este objeto es que se han desarrollado plataformas como Genome Analysis Tool Kit (GATK) (McKenna et al., 2010), que nos ofrecen un ambiente de desarrollo multi-plataforma con un armazón programado en Java(Gosling and McGilton, 1995)organizando la paralelización de todas las herramientas anteriormente nombradas.

### *1.3.3 Identificación de regiones duplicadas del genoma.*

Las zonas codificantes o CDS (CDS, Coding DNA Sequences) presentes en duplicacionesgenómicas se caracterizan por una baja conservación debido a su redundancia funcional, mientras no exista conversión génica(Ohno and others, 1970; Zhang, 2003).Por esto un método utilizado para la identificación de polimorfismos presentes en zonas duplicadas es la categorización de zonas específicas, basado en su nivel de conservación(Andreassen et al., 2010).La caracterización actual del genoma del salmón del Atlántico nos permite aplicar múltiples criterios conocidosde manera innovadora a modo deindicadoresdel nivel de conservación de las regiones del genoma.

#### *1.3.3.1 Indicadores de conservación.*

Para la predicción del nivel de conservación de los intervalos móviles a lo largo de todas las regiones del genoma, nuestro algoritmo utilizó 4 criterios que se detallan a continuación:

*1.3.3.1.1 La razón entre las densidades de polimorfismos presentes entre las zonas codificantes y las no codificantes de un gen.*

Los extremos 5' (río arriba) y 3' (río abajo) no codificantes de los genes, denominadas UTR (UTR, UnTranslated Regions), tienden a verse menos afectadas por fuerzas selectivas que las zonascodificantes, por lo que no solo se espera una mayor densidad de polimorfismos en UTRs versus CDS, sino que también se espera que esta razón disminuya en zonas que presentan parálogos funcionales, ya que al tener sus CDS duplicados homólogos funcionales los hace menos esenciales, presentando una menor conservación(Andreassen et al., 2010).

*1.3.3.1.2 La sumatoria de los efectos delos polimorfismos sobre la expresión del gen.*

Esto considerando que una región del genoma que presenta secuencias parálogas como homólogos funcionales tendráuna menor conservación y por tanto serámás propensa a polimorfismos con efectos más drásticos sobre su expresión y regulación(Cingolani et al., 2012a, 2012b).

*1.3.3.1.3 Desviación de la razón entre transiciones y transversiones (ti:tv).*

La frecuencia de sustituciones de una base en mamíferos se dan entre $1 \times 10^{-9}$ a $5 \times 10^{-9}$ por nucleótido por año (Li et al., 1981; Martínez-Arias et al., 2001), y aunque las frecuencias de cada tipo de sustitución son independientes y dependientes de la especie en cuestión generalmente se encuentran en ordenes de magnitud similares (Collins and Jukes, 1994; Kim et al., 2003; Smith et al., 2001). Es por esto que la probabilidad de la generación espontánea de más de dos alelos (dos sustituciones recurrentes en el mismo sitio del genoma) es muy baja.

Como se nombró anteriormente, en un sistema donde las mutaciones puntuales ocurren de manera aleatoria el radio de transiciones sobre transversiones tiende a 0.5, dado que tenemos lamitad de posibles transiciones (A↔G y C↔T =

4posibilidades) que de transversiones (A↔C, A↔T, G↔C, G↔T = 8 posibilidades)(Strachan and Read, 2004).Sin embargo estas reacciones químicas no tienen las mismas probabilidades de ocurrencia(Vignal et al., 2002; Wondji et al., 2007), incluso cuando son reciprocas. Por ejemplo las deaminaciones de 5'-methyl C a uracilo (C→T o G→A) son el mecanismo más común de mutación y tienen una ocurrencia mucho más alta que el caso inverso, en animales de sangre caliente (Shen et al., 1994; Wyszynski et al., 1994). Es así como en la práctica ti:tv se encontrará sesgada de manera dependiente de cada organismo y los mecanismos moleculares de replicaciónparticulares de estos que puedan favorecer la ocurrencia de algunas de estas reacciones sobre otras (Yang and Yoder, 1999). Por ejemplo en humanos ti:tv es cercano a 2 a lo largo de todo el genoma y en zonas no codificantes se encuentra típicamente cerca de 3(Guo and Jamison, 2005; Zhao and Boerwinkle, 2002), mientras que en especies de peces cercanos a salmón se han observado ti:tv cercanos a 1(Smith et al., 2005).

La ti:tv tenderá a una proporción de equilibrio determinada por los sesgos producidos por los mecanismos moleculares de replicación. Por lo que esperamos que en regiones con mayor taza de ocurrencia de mutaciones aleatorias exista una desviación de éste equilibrio sesgado debida al mayor dinamismo presente en estas regiones.

*1.3.3.1.4 La densidad de candidatos IVs.*

Adicionalmente dado que las zonas parálogas presentan una identidad mayor al 90% (Moghadam et al., 2011), los ensambles obtenidos presentan múltiples lecturas genómicas que difieren de la referencia debido a un mapeo erróneo hacia zonas parálogas (candidatos IVs). Es por esto que la densidad de IVsrepresenta un buen indicador no solo para la identificación de zonas presentes en duplicaciones

genómicas, sino también para la determinación del estado de conservación de estas zonas.

### *1.3.3.2 Modelos de agrupación.*

Luego de la determinación de estos indicadores para cada una de las regiones del genoma nos queda dirimir sobre el estado duplicado y la conservación de las mismas, para lo que requerimos generar un indicador unificado que nos permita agrupar todas aquellas regiones que creemos están en riesgo de tener un mayor dinamismo afectando la conservación de los polimorfismos presentes en ellas.

La agrupación de datos o clusterización se ha convertido en una herramienta estadística de análisis de datos altamente utilizada dentro de la bioinformática, para el reconocimiento de patrones, la minería de datos, el análisis de imágenes, entre otros. La clusterización consiste en el proceso de asignación de uno o varios conjuntos de elementos a un grupo llamado clúster, de modo que los objetos en el mismo clúster sean más similares entre sí (acorde a un conjunto de criterios, parámetros o elementos control de cada grupo objetivo), en comparación con elementos de otros clústeres(Gath and Geva, 1989).

La clusterización consiste en un problema de optimización multi-objetivo que puede ser alcanzada mediante variados algoritmos. El algoritmo apropiado y los parámetros a utilizar para definir la constitución de un clúster dependen de cada conjunto de datos, el objetivo del análisis y sus resultados esperados. En la Tabla II se muestran algunos modelos predictivos de clusterización ampliamente utilizados en bioinformática y sus características.

Los modelospredictivos pueden contrastar su predicción sobre un conjunto de datos no utilizados durante la predicción denominado conjunto de llegada, repitiendo el

proceso predictivo hasta alcanzar un óptimo acercamiento sobre el conjunto de llegada (i.e. sin que se produzca un sobre-ajuste de los datos). A estos modelos se les denomina métodos supervisados.

Sin embargo en los métodos supervisados la generación de una buenapredicción depende dela representatividad delconjunto de llegada, él queestá limitado porla disponibilidad de datos empíricos.En caso de no contar con un buen conjunto de llegada se puede optarpor el uso de un método predictivo no supervisado.

Para la agrupación de nuestras regiones del genoma debemos integrar nuestros indicadores de conservación a los criterios de agrupación. Un métodocapazde integrarcriteriosde manera plástica para dirimir el punto de corte (Tabla II) es el método de clusterización por lógica difusa (Klir and Yuan, 1995),teniendo además la capacidad de ser adaptable para la incorporación de nuevos criterios o modificaciones de los criterios existentes (Ross, 2010).

La clusterización dura asigna a cada dato (vector) un único clúster con un grado de pertenencia igual a uno, asumiendo una clara separación entre clústeres. Este tipo de acercamiento no es capaz de describir de buena forma sub-agrupaciones con límites difusos entre sí.

En el caso del nivel de conservación de las regiones del genoma, los límites de nuestros clústeres no pueden ser determinados a priori, ya que dependen de los niveles de conservación globales del genoma (índices de conservación, ver 1.3.3.1) y su distribución a lo largo de éste. Por esto definiremos un set difuso de valores para cada muestra en base a las distribuciones de los indicadores globales donde el menor grado de conservación tendrá un grado de pertenencia 1 y el menor tendrá un grado de pertenencia 0. El set difuso total es determinado como una función de los conjuntos particulares de cada indicador.

**Tabla II. Modelos estadísticos predictivos para clusterización.**
*(Abdi and Williams, 2010; Byvatov and Schneider, 2003; Klir and Yuan, 1995).*

| MÉTODO | TIPO | Ventajas | Desventajas |
|---|---|---|---|
| Fuzzy Logic (FL) | Lógico | Incorpora Criterio de Experto (i.e. predice basado en el conocimiento actual). | Considerasoloun número limitado de factoresque sabemos están involucrados, puede conllevar a resultados parcializados. |
| Principal Component Analysis (PCA) | No-supervisado | Elimina la parcialidad. | Los resultados de la predicción pueden no tener lógica. |
| Support Vector Machine (SVM) | Supervisado | Considera conjunto de entrenamiento. | Falta de datos para entrenamiento. |

## *1.4 Hipótesis.*

Es posible estimar el estado de conservación de ventanas móviles a lo largo de genomas eucariontes, agrupándolas en segmentos distintivos en base a indicadores de conservación estimablesutilizando lainformación de variantes de secuencia entre regiones parálogas y variantes alélicas entre múltiples individuos. Con el objeto de clasificar todas las variantes obtenidas según el estado de conservación del contexto genómico en el que se encuentre. Obteniendo un conjunto final de variantes de utilidad para la selección genómica.

Siendo este procedimiento de especial utilidad para especies poliploides o pseudo-poliploides donde abundan regiones de baja conservación debidas a la existencia de múltiples parálogos funcionales.

## *1.5 Síntesis.*

El presente trabajo propone un pipeline bioinformático optimizado para la búsqueda de variantes tanto intra-individuo como inter-individuo, utilizando herramientas de última generación: para la obtención, filtrado y caracterización de estos polimorfismos. Adicionalmente nos entrega un acercamiento único e innovador para el descarte *a priori* de variantes no conservadas y la clasificación de las variantes

entre regiones parálogas a partir de un conjunto de marcadores contaminados. Todo esto programado utilizando lenguajes modulares y de código abierto, haciéndolo flexible y capaz de adaptarse a nuevos desafíos.

## *1.6 Objetivos.*

### *1.6.1 Objetivo General:*

El objetivo general del presente proyecto consiste en la generación deun pipeline bioinformático para la obtención de marcadores útiles en estudios de asociación y selección genómica en organismos no diploides o que presenten pseudo-diploidía, utilizando como modelo la especie *Salmo salar* y las bases de datos con información parcial del genoma de 1individuo e información parcial transcripcional de múltiples individuos.

Para alcanzar este objetivo es necesario cumplir con los siguientes objetivos específicos.

### *1.6.2 Objetivos Específicos:*

#### *1.6.2.1 Determinación de variaciones Intra-individuo.*

Consiste en la determinación y obtención del conjunto de variantes entre secuencias genómicas parálogas de un individuo doble haploide (PSVs).

#### *1.6.2.2 Determinación de variaciones Inter-individuo.*

Consiste en la determinación y obtención del conjunto de variantes entre las mismas secuencias genómicas entre múltiples individuos (SNPs/DIPs/MSVs). Conteniendo información falsa debida a las variaciones entre secuencias parálogas del genoma (PSVs), además de variantes poco conservadas sin utilidad para la selección asistida para la cruza.

#### *1.6.2.3 Predicción de zonas duplicadas del genoma del salmón.*

Determinarun indicador unificado del nivel de conservación de múltiples ventanas dentrodel genoma y clasificándolas acorde a éste.

#### *1.6.2.4 Validación de zonas duplicadas predichas contra True-SNPs empíricos.*

Verificar la consistencia de nuestro método predictivo por comparación contra un set de SNPs validados empíricamente.

#### *1.6.2.5 Generación de un conjunto de True-SNPs del genoma completo del salmón.*

Conjunto de SNPs eficientes para su uso en selección genómica.

# 2 MATERIALES Y METODOS

## 2.1 MATERIALES

### 2.1.1 Secuencias:

#### 2.1.1.1 Referencia:

Todos los análisis de la presente tesis se realizaron sobreun ensamble de secuencias extraídas de un individuo *Salmo salar* (salmón del Atlántico) hembra doble haploide conocido como "Sally".Se utilizó el ensamble público denominado AGKD01 disponible en NCBI (Davidson et al., 2010), generado por el consorcio internacional "International Cooperation to Sequence the Atlantic Salmon Genome" (ICSASG)utilizando el ensamblador *de novo*"whole-genome shotgun"ABySS (v1.2.7)(Simpson et al., 2009) para las lecturas Illumina y Arachne (v3.9) (Batzoglou et al., 2002)combinando los contigs mayores a 500bp con PCAP (Huang et al., 2003).Este ensamble presenta un N50 de 9,3Kb,longitud total de 2.43Gb;contabilizando por un 76% del genoma del salmón.

Para su pre-procesamiento se utilizaron: las bases de datos de secuencias de elementos repetitivos específicas de salmónidos de cGRASP (Ng et al., 2005) versión 1.6,disponible públicamente (http://web.uvic.ca/grasp); y las bases de datos de elementos repetitivos estándar proporcionadas por RepBase (Jurka et al., 2005) descargadas en Abril del 2012.

#### 2.1.1.2 Intra-individuo:

Para la determinación de las variantes intra-individuo seutilizaron 238millones de lecturas"paired-end" Illumina (141.728.293 lecturas de tamaño de inserto de 600 bp

y 96.683.324 lecturas de tamaño de inserto de 300 bp),provenientes de gDNA de músculo deSally (hembra *Salmo salar* doble haploide), obtenidas por ftp a través del ICSASG y 28 millonesde secuencias Sanger públicas (28.208.160 lecturas),provenientes de la base de datos TraceDB de NCBI(ftp://ftp.ncbi.nlm.nih.gov/ pub/TraceDB), pertenecientes a Sally.

#### *2.1.1.3 Inter-individuo:*

Para la determinación de las variantes inter-individuo se utilizaron:495 K (495.211) de lecturas Expressed Sequence Tags (ESTs) sin calidades, provenientesde múltiples individuos *Salmo salar,* descargadas de la base de datos pública dbEST(Boguski et al., 1993), con un largo promedio de 617 bp;129 K (128.640) desecuencias mate-pairde *Salmo salar*provenientes de una librería Bac End Sequence (BES) de trabajo previo realizado por Thorsen con DNA genómico de un macho de una población Noruega (Thorsen et al., 2005), con un largo promedio de 900 bp y solicitada vía ftp al PhD. Willie Davidson en Simon Fraser University, Canadá; y 143M (142.645.370) de lecturaspaired-endIllumina de largopromedio de 100 bpprovenientes de 16 individuos de criaderos chilenos,proporcionadas por la empresa AQUAINNOVO S.A..

#### *2.1.1.4 Predicción de estructura de genes:*

Para la predicción de las estructuras génicas se utilizaron los EST anteriormente nombrados (ver 2.1.1.3) en conjunto con lasproteínas caracterizadas de *Salmo salar* provenientes de la base de datos pública proteindb de NCBI(Benson et al., 1997; Pruitt et al., 2007).

#### *2.1.1.5 Validación in sillico:*

Para la validación se utilizó un conjunto de 5.768 SNPs validados para *Salmo salar*en una población noruega (29 cromosomas) (Lien et al., 2011), publicadosen la base de datos dbSNP(Sherry et al., 2001).

### *2.1.2 Hardware:*

Para llevar a cabo estos análisis se utilizaron: un clúster IBM de 528 nodospúblicos de 2,67GHz Intel® Xeon® CPU X5550, 24GB de memoria por nodo público y 8TB de almacenamiento, conectado por Infiniband QDR 100% non-blocking, denominado "levque" perteneciente al "High Performance Computing Laboratory" (HPCLAB) del Centro de Modelamiento Matemático (CMM) de la Universidad de Chile; un servidor de 16 cores de 2,9GHz Intel® Xeon® con 120GB de RAM y 7,5TB de almacenamiento denominado "bio-5" perteneciente a Mathomics, CMM, Universidad de Chile; y un servidor web "espora" con 4 cores de 2,8GHz Intel® Xeon® y 8GB de RAM, perteneciente a Mathomics, CMM, Universidad de Chile.

### *2.1.3 Software:*

Para el desarrollo del protocolo se eligieron las herramientas y formatos según los criterios enumerados en la Tabla III.

### *2.1.4 Financiamiento:*

El proyecto fue financiado por el Programa Fondo de Investigación Avanzado en Áreas Prioritarias (FONDAP), la empresa AQUAINNOVOS.A., el Programa de Tesis de Postgrado en la Industria CONICYT 2011 y el Centro de Regulación Genómica (CRG) de la Universidad de Chile.

**Tabla III. Herramientas y formatos utilizados para el desarrollo del protocolo sugerido.**

*Las herramientas estándar son referenciadas en la sección métodos(Blanca et al., 2011; Cingolani et al., 2012a, 2012b; Danecek et al., 2011; Dean and Ghemawat, 2008; Ihaka and Gentleman, 1996; Lam et al., 2012; Langmead and Salzberg, 2012; Li and Durbin, 2009; Li et al., 2009; McKenna et al., 2010; Schwab et al., 2000; Stajich et al., 2002; Stallman and McGrath, 1988; Wu and Nacu, 2010; Wu and Watanabe, 2005).*

| Herramienta | Descripción | Ventajas |
|---|---|---|
| clean_reads | Limpieza, recorte y filtradode lecturas de NGSa travéz de:remoción de adaptadores y vectores; recortado de regiones de baja complejidad o baja calidad; y filtrado de secuencias de largo insuficiente. | Multithreading[11].Flexible, puede interpretar lecturas Sanger, 454, Illumina and SOLiD[11].Gran incremento en exactitud para BWA and GMAP mapping[11]. |
| BWA | Alineamiento y mapeo de lecturas NGS con baja divergencia a referencias de genoma completo.<br>Usa Burrows-Wheeler transform (BWT)[7]. | Flexible, puede mapear lecturas cortas Illumina de 100 b así como lecturas largas de 70 b a 1 Mb. Actualizado. |
| GMAP[10] | Mapeo y alineamiento de lecturas transcriptómicas a un genoma de referencia[10].<br>Usa un algoritmo basada en hashing. | Menor consumo de memoria (lectura directa)[10].Aumenta mapeo y predicción de exones[10].Menor número de falsos positivos a mayor velocidad[10].Mejorprecisión para lecturas de gran longitud[10].Generación de una estructura génica fiel[10].Tolerante a SNPs[10]. |
| Picard Tools | Herramientas Javapara la manipulación de archivos SAM/BAM[8]. | Combinalos encabezados de múltiples archivos. |
| GATK[10] | Marco de red para programación estructurada diseñado para facilitar el desarrollo de herramientas de análisis robustas y eficientes para NGS[10], usando MapReduce[6]. | Designación de variantes de mayor especificidad sin sacrificar sensibilidad[15]. |
| SnpEff[13,14] | Programa computacional para la rápida determinación, anotación, clasificación y categorización delos efectos de variantessobrela transcripción y traducción en secuencias de genoma completo[13]. | Permite fusionar y comparar variantes. Mejora los criterios de selección de variantes. |
| **Formato** | **Descripción** | **Ventajas** |
| SAM/BAM[8] | Formato de alineamiento de secuencia genérico para el almacenamiento de alineamientos de lecturas contra una referencia. | Formato flexible, compacto, binario (BAM), eficiente, de rápido accesoparadeterminación de variantes[8]. Soportalecturas NGS largas (< 128 Mbp)[8]. |
| VCF[12] | Formato genéricopara el almacenamiento de polimorfismos de DNA, como SNPs, DIPs yvariantes estructurales. Almacena la posición en conjunto con abundantes anotaciones[12]. | Formato compact, rápida obtención de datos (a través de indexación), puede acotar el análisis a rangos de datos determinados[12]. |
| **Lenguaje** | **Descripción** | **Ventajas** |
| Perl/Bioperl[4] | Proyecto de colaboración internacional multidisciplinariode código abierto, parael desarrollo de librerías modulares en lenguaje Perl para la administración y manipulación de información de las ciencias biológicas[4]. | Reduce el tiempo de desarrollo, a un nivel de desempeño eficiente[4].Flexible ycapaz de interoperarcon otros lenguajes (Python and Java)[4].Lenguaje de más alto código comparado con Biopython[4]. |
| R[2] | Lenguaje para el análisis y gráfico de datos[2]. | Portabilidad, eficiencia de computo, mejor manejo de memoria y alcance[2]. |

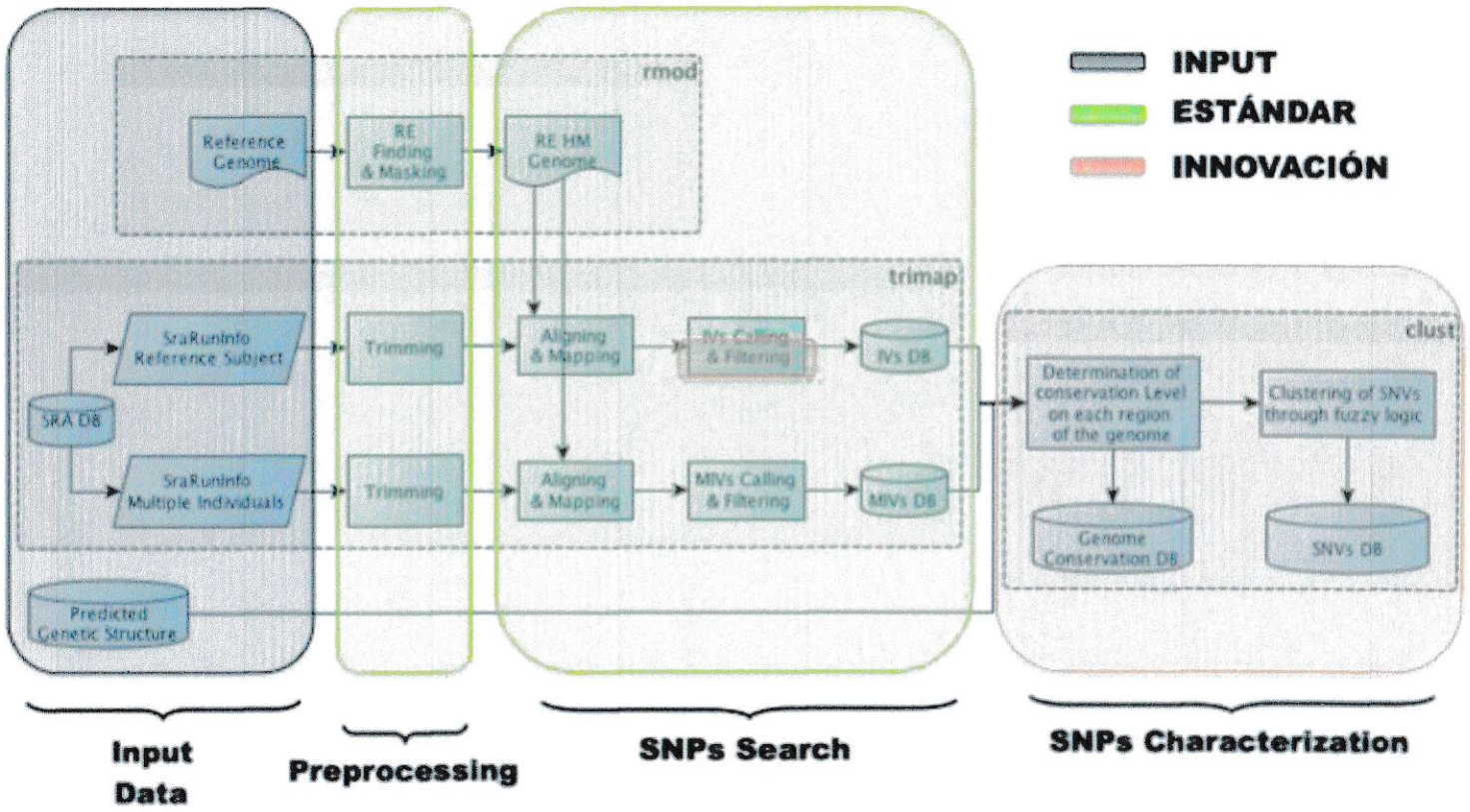



**Figura 11. Pipeline estándar del Kit de búsqueda y caracterización de SNPs SNP-SACK.**
*El pipelineestácompuesto por 4procesos principales(Input Data): Entrada de datos mínimos, requiere un genoma de referencia y secuencias de múltiples individuos; (Preprocessing): Involucra el trimming, descarte y enmascaramiento, defragmentos sin utilidad y secuenciasdel conjunto inicial. (SNPs Search): Búsqueda de variantes entre parálogos (IVs) y entre múltiples individuos (MIVs). (SNPs Characterization): Predicción de regiones"duplicadas" basadoen "indicadoresconservación".*

## 2.2 MÉTODOS.

Para la obtención y validación del conjunto de marcadores segregantes conservados del genoma del salmón del Atlántico se propone el protocolo mostrado en el diagrama de laFigura 11.

Los scripts utilizados por el protocolo fueron programados en Perl (BioPerl), Bash o R y el protocolo completo fue llevado a un script automatizado programado utilizando Perl.

Las herramientas y paquetes informáticos existentes que se utilizaron y sus criterios de elección se especifican en la Tabla III.

### 2.2.1 *Pre-procesamiento del ensamble.*

Se descartó el conjunto de elementos repetitivos predichos en el genoma del salmón mediante la herramienta RepeatMasker (Smit et al., 2004; Tarailo-Graovac and Chen, 2009), utilizando las bases de datos de elementos repetitivos específicas de salmónidos (Ng et al., 2005) y las proporcionadas por RepBase (Jurka et al., 2005).

Se pre-filtraron los contigs de largo menor a 1 Kb, para evitar contigs que contengan errores de ensamble(Huang et al., 2011).

### 2.2.2 *Obtención de Variaciones.*

Todas las lecturas utilizadas para la obtención de variaciones fueron pre-filtradas y recortadas utilizando el pipeline automatizado clean_reads desarrollado por COMAV Bioinformatics (Blanca et al., 2011), eliminando regiones de baja calidad, adaptadores de secuenciamiento, vectores de las librerías y zonas de baja complejidad (Tabla III), y luego excluyendo las lecturas de largo insuficiente en pb (< 100 bp).

#### *2.2.2.1 Mapeo de lecturas Intra-individuo.*

Para detectar la presencia de polimorfismos en los contigs de nuestro ensamble, se mapearon:el conjunto de 13millonesde secuencias genómicas Sanger y 32 millones de secuencias genómicas Illuminaprovenientes de librerías públicas utilizando BWA (Li and Durbin, 2009, 2010). Los resultados se almacenaron en formato de secuencia de mapeo/alineamiento binario (Binary Sequence Alignment/Map, BAM) (Li et al., 2009).

#### *2.2.2.2 Mapeo de lecturas Inter-individuo.*

Laslecturasprovenientes de EST se mapearon utilizando GMAP (Wu and Nacu, 2010; Wu and Watanabe, 2005). Las lecturas provenientes de DNA genómico (BES e Illumina), se mapearonmediante BWA(Li and Durbin, 2009, 2010). Los resultados se filtraron por cobertura (DIPs) e identidad (SNPs) mayor al 98% evitando zonas de baja calidad caracterizadas por una alta densidad de variaciones, los resultados se almacenaron en formato BAM(Li et al., 2009).

#### *2.2.2.3 Filtrado de variantes.*

Los polimorfismos candidatos intra-individuo (IVs) e inter-individuo (MIVs) fueron predichos y filtrados utilizando la suite GATK (McKenna et al., 2010), utilizando el filtro de tipo "base alignment quality" (BAQ) (Li, 2011), y los resultados fueron almacenados en formato Variant Call Format (VCF), de 1000 Genomes Project (Danecek et al., 2011). Finalmente las variantes restantes fueron refiltradas según los criterios enumerados en laTabla IV.Se utlizaron 2 filtros innovadores específicos para IVs, considerando que el número máximo de parálogos posibles es de 2 elevado al número de WGDR sufridos por la especie y un último filtro para excluir regiones donde se encontraron MNP en IVs desde el conjunto de MIVs obtenidos, evitando variantes debidas a errores de mapeo.

**Tabla IV. Criterios de filtrado de variantes.**
*Criterios de descartepara la exclusión de variantes no utilizables previo a la determinación de los índices de conservación.*

| Criterio | IVs | MIVs | Objetivo |
|---|---|---|---|
| $Qual \geq 30$ | Aplica | Aplica | Incrementarla confianzaen lasdesignaciones de variantes. |
| $Depth \geq 4$ | Aplica | Aplica | Descartedeerrores de secuenciamiento. |
| $Var \geq \left(\frac{1}{2}\right)^R$ | Aplica | NA | Descarte de errores de secuenciamiento. |
| $Var \geq 0{,}1\ \%$ | NA | Aplica | Descarte de variantes inservibles para el genotipado. |
| $MapQual \geq 10$ | Aplica | Aplica | Descarte de errores de mapeo. |
| A 10 bp min. de gap | Aplica | Aplica | Descarte de errores de ensamble. |
| A 10 bp min. de poli-A. | Aplica | Aplica | Descarte de errores de secuenciamiento. |
| MNP | Aplica | NA | Descarte de errores de mapeo. |
| $No\ alelos \leq 2$ | NA | Aplica | Descarte de SNPs falsos. |
| $No\ alelos \leq 2^R$ | Applica | NA | Descarte de IVs falsos. |

### *2.2.3 Predicción de estructuras génicas.*

Las estructuras génicas se predijeron utilizando los resultados del mapeo de las secuencias ESTs y proteínaspúblicas de *Salmo salar,*contra el ensamble de referencia utilizado (AGDK01), mediante un acercamiento *ab initio*(Di Genova et al., 2011).

### *2.2.4 Clusterización.*

Mediante el paquete bioinformático SnpEff (Cingolani et al., 2012a, 2012b) (incluído en GATK)se calculó el efecto de las variantes polimórficas, a partir de las estructuras génicas predichas y los conjuntos de variantes(incluyendo tanto variantes Inter como Intra-individuo). Adicionalmente se obtuvo un archivo resumende la organización de las estructuras génicas del genoma.

Al Pipeline junto con el conjunto de scripts utilizados y desarrollados para la clusterización y filtrado de variantes según su nivel de conservación se le denominó

SNP Search and Characterization Kit (SNP-SACK).Los indicadores de conservación (Tabla V) de los genes predichos e intervalos de ventanas móviles a lo largo del genoma, se calcularon mediante la cruza del archivo resumen de las estructuras génicas del genoma y los conjuntos de variantes;ésto utilizando un script llamadoeff_join.pl (SNP-SACK --clust). Los tamaños mínimo y máximo del intervalo se calcularon en base al N75 y máximo de la distribución de tamaño de los genes predichos.

Los archivos de salida producidos son:

1) Un archivo único en formato VCF (Danecek et al., 2011) que incluye las variantes producto del cruce entre los candidatos IVs y los candidatos MIVs.

2) Una lista de las estructuras génicas predichas y sus tamaños eíndices de conservación correspondientes.

3) Una lista de los intervalos generados y sus índices de conservación asociados.

Estos resultados son luego procesados por un script desarrollado enlenguaje R (Ihaka and Gentleman, 1996), llamadotab.R (SNP-SACK), el cual realiza un análisis predictivo en tres pasos:

1) Determinacióndel umbral de corte para los set difusos de las estructuras génicas, utilizando los índices mínimos del conjunto resultante. Determinado como el vertice de la distribución asintótica de set difuzos unificados para las estructuras génicas utilizadas.

$$\mu_{unif} = \mu_{eff} \cdot \sqrt{\mu_{\delta IVs}} \cdot \frac{\mu_{\delta Var_{EXON}}}{\delta Var_{UTR}} \cdot \sqrt{\mu_{|ti:tv_n - \overline{ti:tv}|}}$$

$$th: (\mu_n - \mu_{n-1})N = 1$$

**Figura 12. Ecuaciones del set difuso unificado y determinación del punto de corte (th).** *Se concentró para ambos criterios aplicables a todo el rango, El th se determinó como el vértice de la distribución asintótica del histograma de los sets difusos.*

2) Determinación de la función predictiva del set difuso a partir de los 2 índices de conservación aplicables al genoma completo (ver Tabla V); 3) cruza y descartede marcadores candidatospresentes en intervalos con una distribución deíndicespor debajo del umbral (th) calculado.

$$\mu' = \sqrt{\mu_{\delta IVs}} \cdot \sqrt{\mu_{|ti:tv_n - \overline{ti:tv}|}}$$

$$\mu_{unif} = \mu' \pm f(\mu')$$

**Figura 13. Ecuaciones del sub set difuso y determinación del rango difuso unificado.** *El sub set difuso se determina utilizando ambos criterios aplicables concentrados. El rango de criterios difusos unificados posibles se determina como función del sub set difuso.*

Los indicadores de conservación aplicables a los intervalos de ventanas móviles (Tabla V) fueron diluidos en el cálculo del conjunto difuso de las estructuras génicas con el objetivo de mejorar la sensibilidad de nuestra predicción. Se utilizó un criterio de corte para el rango difuso de una pertenencia equivalente al 0,75 para mejorar la especificidad de nuestro método.

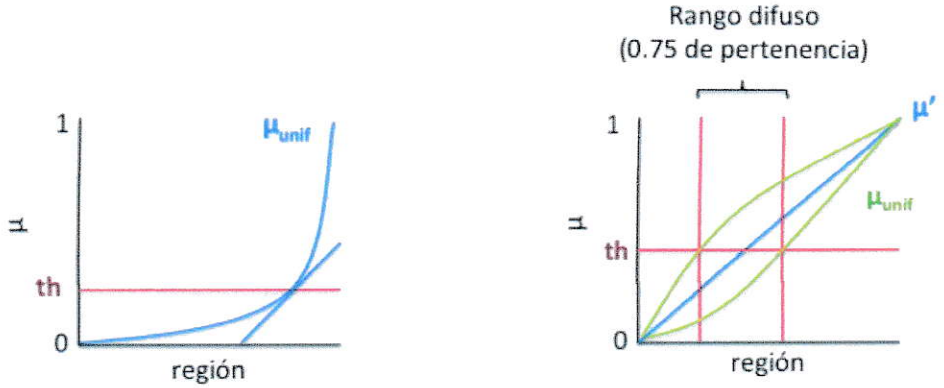



**Figura 14. Representación gráfica de una distribución de sets difuzos en un conjunto de observaciones.**
a) Una distribución del set difuso marcando el punto de corte (th) obtenido según la ecuación anterior. b) Gráfico logarítmico representativo del rango difuso unificado (verde)estimadoen función de un sub set difuso (azul), generando un rango difuso. En nuestro caso se determinó un 0,75 de pertenencia como criterio de corte para el rango difuso.

Los resultados finalesfueron inspeccionados visualmente e *in sillico*,según criterios de literatura (1 variante cada 650pb)(Moen et al., 2008).

**Tabla V. Indicadores de conservación utilizados como criterios de corte.**
*(Andreassen et al., 2010; Cingolani et al., 2012b, 2012b; Fredman et al., 2004; Strachan and Read, 2004).*

| Criterios | Cálculo |
|---|---|
| *Effect*[4,5] | Sumatoria de los efectos valorados calculados por SnpEff. |
| $\delta IVs^1$ | DensidaddeCandidatos PSVs/MSVs por longitud de secuencia en bases. |
| $\frac{\delta Var_{EXON}}{\delta Var_{UTR}}^3$ | Densidad de variaciones predichas acumuladas totales de regiones exónicas (de mayor conservación) sobreregiones no-expresadas (de menor conservación). |
| $\lvert ti: tv_x - ti: tv_\mu \rvert^2$ | Taza de transiciones sobre transversionesdelos SNPs candidatos. |

### 2.2.5 *Validación in sillico.*

Los resultados se validaron utilizando un conjunto de marcadores(5.768) caracterizados y validados provenientes de ESTs y genes asociados (Lien et al., 2011), publicados en NCBI.

Los marcadores se mapearon utilizando MegaBLAST(Zhang et al., 2000) y fueron filtrados por identidad mayor al 98%.

# 3 RESULTADOS

En esta sección se resumen los resultados obtenidos al aplicar el pipeline bioinformático propuesto y desarrollado en esta tesis (SNP-SACK) que se muestra en la Figura 11 y cuya metodología sedescribe en la sección 2.2.

## *3.1 Pre-procesamiento del ensamble.*

Luego de la obtención del ensamble del genoma de *Salmo salar*Sally AGKD01, y previo a su uso como genoma de referencia se procesó para la búsqueda de elementos repetitivos (ER), con lo que se predijeron 4,12 M de ERcon un largo total de 1,36 Mb, enmascarando un 59 % del ensamble utilizado. La distribución de estos ER a lo largo del ensamble se grafica en la Figura 15.

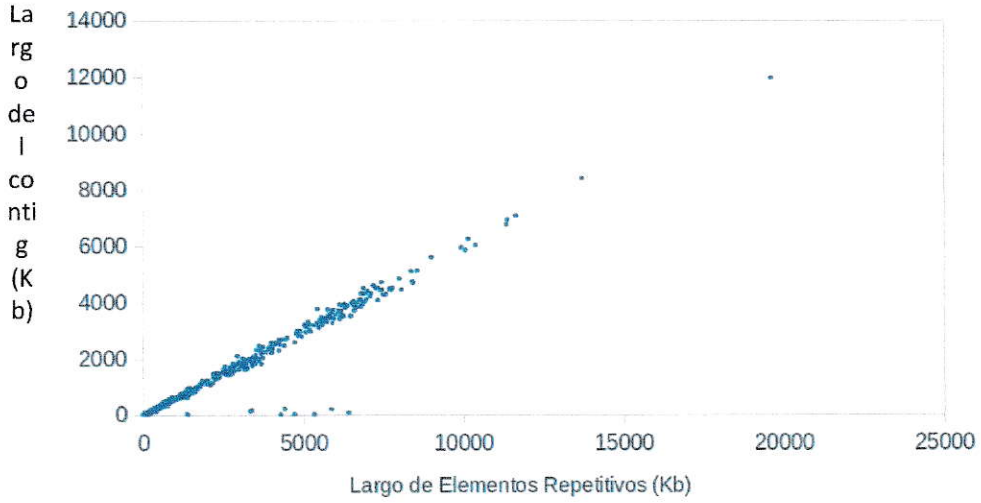



**Figura 15. Distribución de largos de elementos repetitivos con respecto a los largos de los contigs.**
*La figura muestra una tendencia lineal en la relación de largo de elementos repetitivos presentes en un contig y la longitud de este.*

Con el objeto de mejorar la calidad del genoma de referencia a utilizar mediante el descartede contigsformados por errores de ensamble se filtraron los contigsde largos menores a 1 Kb obteniendo un largo total remanente de 2,04 Mb (contabilizando para una representatividad del ~67% del genoma del salmón del Atlántico), lo que fue utilizado como genoma de referencia para la obtención de marcadores candidatos en *Salmo salar*.

## *3.2 Obtención de marcadores candidatos.*

Previo a la obtención de variantes tanto inter como intra-individuo se realizó un recorte o "trimming" de las lecturas. Los ajustes y configuracióndel programa clean_read de ngs_backbone pipeline(Blanca et al., 2011)se adecuaron al origen de las lecturas, como se especifica en los manuales y opciones de ayuda del paquete. Luego de la limpieza de las lecturas se procedió al mapeo de lecturas dependiendo de su origen como se especifica en materiales y métodos (ver 2.2.2).

En la Tabla VI se resumen la tecnología de secuenciamiento utilizada para la obtención o tipo de base de datos de las lecturas utilizadas, su naturaleza, la fuente desde la que se obtuvo la secuencia, las fechas de descarga de las secuencias públicas y el número total de lecturas iniciales obtenidas. Adicionalmente se resumen los resultados obtenidos (i.e. porcentaje de lecturas remanentes):luegodel trimming mediante "clean_read" y filtrado de lecturas de largo insuficiente; y el mapeo mediante BWA o GMAP contra el genoma de referencia de cada conjunto de lecturas utilizado.

Los mapeos de múltiples individuos fueron filtrados para una coberturae identidadmayor al 98% eliminando posibles mapeos erróneos. Para los mapeos de las lecturas provenientes de "Sally" (Intra-individuo) se filtraron los mejores

16alineamientos, tomando en cuenta el máximo de regiones parálogas posibles dado el número de rondas de duplicación sufridas por la familia de los salmónidos.

**Tabla VI. Resumen de filtrado y alineamientos obtenidos para los conjuntos de lecturas utilizadas.**
*Los porcentajes de lecturas filtradas y mapeadas se calcularon con respecto al número de lecturas iniciales utilizadas. La cobertura se calculó considerando el tamaño estimado del genoma completo del salmón del Atlántico (~3 x 10 $^{9}$ bp).*

| Obtención de secuencias | ORIGEN | FUENTE | FECHA | Lecturas iniciales | Lecturas filtradas | Lecturas mapeadas | Cobertura |
|---|---|---|---|---|---|---|---|
| **INTER-IND** | | | | | | | |
| EST db | EST | NCBIESTdb | 04/26/12 | 495.211 | 99,81% | 97,23% | 0,11X |
| Sanger mape-pair | gDNA BES | (Thorsen et al., 2005) | - | 128.640 | 95,54% | 76,24% | 0,07X |
| Illumina Paired-end | gDNA | Aquainnovo S.A. | - | 142.645.370 | 99,36% | 62,38% | 12,31X |
| **INTRA-IND** | | | | | | | |
| Illumina Paired-end (Sally) | gDNA muscle | ICSASG | - | 238.411.617 | 97,90% | 95,00% | 19,97X |
| Sanger (Sally) | gDNA | TraceDB | - | 28.208.160 | 91,24% | 95,29% | 2,35X |
| Protein db | - | Salmondb/NCBI proteindb | 04/26/12 | 24.423 | - | 76,23% | - |
| SNP db | All | Salmondb/NCBI snpdb | 04/26/12 | 88.932 | - | 91,44% | - |

A partir de los archivos BAM (Li et al., 2009)obtenidos por mapeo de laslecturas enumeradas en la Tabla VI, se procedió a la determinación de variantes entre las secuencias,obteniendo archivos en formato VCF que luego fueron filtrados aplicando los criterios enumerados en la Tabla IV. Se obtuvieron 35.038 marcadores candidatos a partir de múltiples individuos (MIVs), de los cuales 3.340 fueron DIPsy 4.857.584PSVs/MSVs candidatos provenientes del mapeo de secuencias de Sally (IVs).

## *3.3 Predicción de estructuras génicas.*

Al usar un predictor *ab initio* en conjunto con herramientas estándar de predicción de estructuras génicas se obtuvieron 24.062 genes candidatos de los cuales: 7.869 constituyen genes candidatos completos (i.e. comprendiendo desde 5'-UTR hasta 3'-UTR); 8.162 son genes candidatos parciales desde 5'-UTR; 3.792 son genes candidatos parciales desde 3'-UTR; y 4.239 genes candidatos parciales internos.

## *3.4 Caracterización de polimorfismos.*

Consecutivamente los conjuntos de polimorfismos candidatos obtenidos en 3.2, fueron caracterizados como se especifica en 2.2.4mediante las estructuras génicas predichas por nuestro laboratorio, determinando la distribución relativa de estos marcadores y sus efectos sobre la expresión y regulación de la expresión en el genoma.

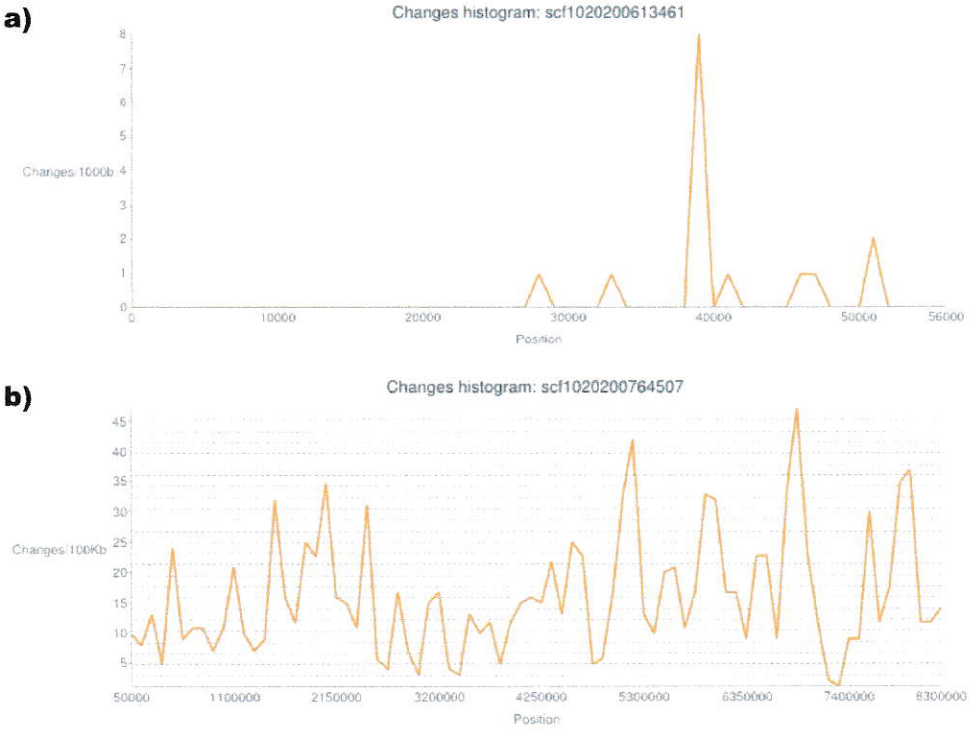



**Figura 16.Efecto de las duplicaciones en la densidad de marcadores.**
*Dos histogramas representativos de las frecuencias de MIVs a lo largo del contig. a) Muestra un contig regular con densidades esperadas según propuesto por Hayes (~ 1 SNP cada 650bp). b) Un contigmostrando sobre 1 SNP cada 50 bp (Hayes and Goddard, 2010), causado probablemente por errores de mapeo debido a regiones duplicadas.*

El archivo resumen obtenido del análisis de los candidatos por SnpEff muestra regiones del genoma con histogramas dedensidades de marcadores que se alejan de lo esperado en salmónidos (una variante cada 650 pb, Figura 16a ), y se asemejan a lo proyectadopara regiones genómicas que presentan parálogos (una variante cada 50 pb, Figura 16b) (Hayes et al., 2001).

## 3.5 Clusterización.

Luego de la caracterización de los polimorfismos se procedió al cálculode los índices de conservación para los genes candidatos,enumerados en la Tabla V, y el set difuso resultante de la distribución de estos cuatro, acorde a lo descrito en la sección 2.2.4 del presente trabajo de tesis (diluyendo los sets difusos de ti:tv y $\delta$lVs).Todo lo anterior utilizando scripts propios desarrollados en R (tab.R de SNP-SACK). Luego se recalcularon los sets difusos para los genes candidatos utilizando solo los indicadores diluidos, calculando la diferencia entre ambos sets con lo que se obtuvo la correlación representada en laFigura 17.

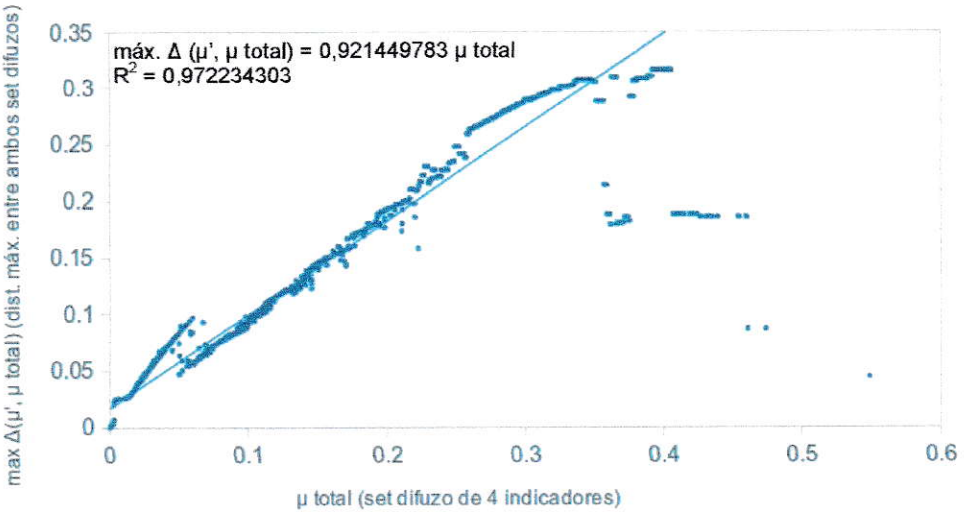
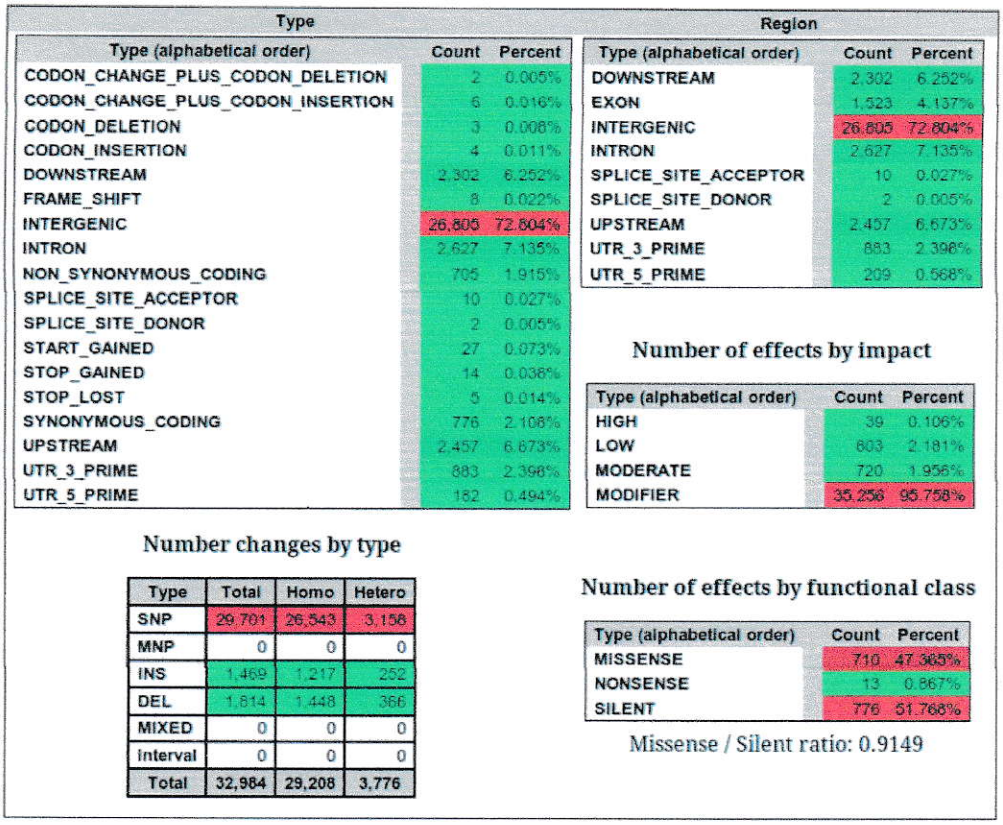



**Figura 17.Correlación entre set difuso total y de intervalo, en regiones codificantes.** *μ total representa el set difuso calculado utilizando los 4 indicadores de conservación para cada región codificante predicha. μ' representa el set difuso de las mismas regiones codificantes calculado a partir de los 2 indicadores de conservación aplicables a todo intervalo de ventana a lo largo del genoma. La correlación (0,968) muestra tendencia a la linealidaden todo el rango por debajo de un valor de difusidad del 40%.*

En la figura anterior se observa el ajuste a una función lineal de la correlación entre el set difuso total (calculado a partir de los 4 indicadores, verTabla V) y su distancia al set difuso aplicable a los intervalos de ventanas móviles (calculado a partir de los 2 indicadores aplicables, verTabla V). La función de esta recta nos permite acotar el rango de valores posibles para los sets difusos de los intervalos de ventanas

móviles, acorde con los valores de los 2 indicadores aplicables a estos (ti:tv y $\delta$IVs).Podemos estimar un valor de corte para la clusterización correspondiente al vértice de la gráfica asintótica de la distribución de los valores de los sets difusos para los genes predichos. La media de la distribución de sets difusos obtenida fue de 0,9314.

Tomando en consideración estos valores se generaron rangos de sets difusos posibles para cada intervalo de ventana. Con esto se generó una clusterización de intervalos de ventana donde se identificaron 3 grupos difusos: 1) de alta conservación o "blancos" 2) de moderada conservación o "grises" (rango difuso) 3) y de escaza conservación o "negros". Los rangos negros fueron excluidos y los rangos "grises" fueron excluidos considerando un rango de pertenencia del 0,75.A través de este procedimiento y a partir del conjunto original de 31.698SNPs, 3.340 DIPs y 4.857.584IVs,se obtuvo un conjunto de 29.701 SNPs y 2.973 DIPs en regiones de alta conservación (6,30% y 10,99% descartados respectivamente, 6.75 % total); mientras que el 36,75 % de los IVs fueron descartados. La distribución y efectos del conjunto final de Variantes se muestran en la Figura 18.



Type

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| CODON_CHANGE_PLUS_CODON_DELETION | 2 | 0.005% |
| CODON_CHANGE_PLUS_CODON_INSERTION | 6 | 0.016% |
| CODON_DELETION | 3 | 0.008% |
| CODON_INSERTION | 4 | 0.011% |
| DOWNSTREAM | 2,302 | 6.252% |
| FRAME_SHIFT | 8 | 0.022% |
| INTERGENIC | 26,805 | 72.804% |
| INTRON | 2,627 | 7.135% |
| NON_SYNONYMOUS_CODING | 705 | 1.915% |
| SPLICE_SITE_ACCEPTOR | 10 | 0.027% |
| SPLICE_SITE_DONOR | 2 | 0.005% |
| START_GAINED | 27 | 0.073% |
| STOP_GAINED | 14 | 0.038% |
| STOP_LOST | 5 | 0.014% |
| SYNONYMOUS_CODING | 776 | 2.108% |
| UPSTREAM | 2,457 | 6.673% |
| UTR_3_PRIME | 883 | 2.398% |
| UTR_5_PRIME | 182 | 0.494% |

Region

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| DOWNSTREAM | 2,302 | 6.252% |
| EXON | 1,523 | 4.137% |
| INTERGENIC | 26,805 | 72.804% |
| INTRON | 2,627 | 7.135% |
| SPLICE_SITE_ACCEPTOR | 10 | 0.027% |
| SPLICE_SITE_DONOR | 2 | 0.005% |
| UPSTREAM | 2,457 | 6.673% |
| UTR_3_PRIME | 883 | 2.398% |
| UTR_5_PRIME | 209 | 0.568% |

Number of effects by impact

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| HIGH | 39 | 0.106% |
| LOW | 803 | 2.181% |
| MODERATE | 720 | 1.956% |
| MODIFIER | 35,256 | 95.758% |

Number changes by type

| Type | Total | Homo | Hetero |
|---|---|---|---|
| SNP | 29,701 | 26,543 | 3,158 |
| MNP | 0 | 0 | 0 |
| INS | 1,469 | 1,217 | 252 |
| DEL | 1,814 | 1,448 | 366 |
| MIXED | 0 | 0 | 0 |
| Interval | 0 | 0 | 0 |
| Total | 32,984 | 29,208 | 3,776 |

Number of effects by functional class

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| MISSENSE | 710 | 47.365% |
| NONSENSE | 13 | 0.867% |
| SILENT | 776 | 51.768% |

Missense / Silent ratio: 0.9149

**Figura 18. SNPs finales y sus efectos por tipo y región.**
*La tabla muestra los resultados del análisis del conjunto final de SNPs luego de la clusterización y descarte de regiones con baja conservación.*


**Figura 19.Corrección de la densidad de marcadores en zonas duplicadas por SNP-SACK.**
*Dos histogramas representativosde la frecuencia de variaciones según su distribución a lo largo del mismo contig antes y después de la aplicación de SNP-SACK. a) Representa un contigcon regiones de alta densidad de variantes (sobre 1 SNP cada 50 bp), distribuidas de manera irregular. b) Representa el mismo contig luego del tratamiento de este con SNP-SACK donde se eliminaron los marcadores de menor conservación dejando un marcador altamente conservado para su uso en selección genómica.*

La Figura 19 nos muestra dos histogramas de densidades de variantes a lo largo de un contig, antes (Figura 19a) y después (Figura 19b) del filtrado con SNP-SACK.

La Figura 19a muestra densidades cercanas a 1 SNP cada 50 bp, predichas para zonas duplicadas en salmón(Hayes and Goddard, 2010),mientras que en la Figura 19bse observa una clara reducción luego del filtrado mediante SNP-SACK.

Las 30 K variantes obtenidas marcan eficientemente para un 22,23 % del total del ensamble (627 Mb).

## *3.6 Validación in silico.*

Del conjunto original de 5.768 marcadores públicos (Lien et al., 2011), obtenidos desde dbSNP (Sherry et al., 2001) de NCBI, el 65,25% mapearon a nuestro genoma de referencia con alta identidad (sobre el 98%), 2.984 SNPs, 701 MSV-3s y 67 MSV-5s.

De las variantes públicas mapeadas:96,05% de los SNPs,86,16% de los MSV-3s y 83,58% de los MSV-5sse ubicaron en regiones de "alta conservación" según nuestro método predictivo, evidenciando la especificidad del método.

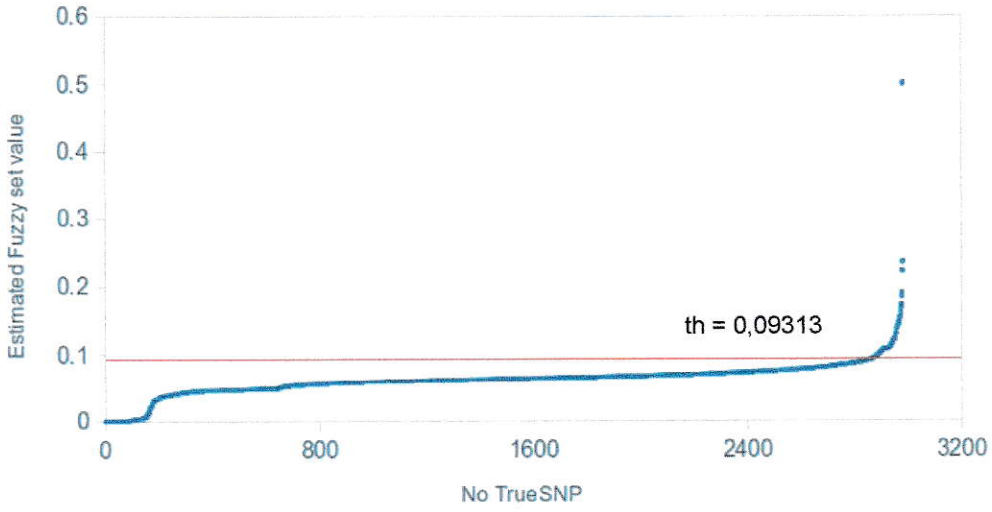



**Figura 20. Distribución de sets difusos estimados para el conjunto de True-SNPs.**
*La distribución delamuestra se ajusta a una curva logarítmica con asíntota cerca del 4%, coincidente con el valor de threshold (th) estimado mediante lógica difusa (0,09313).*

La figura anterior muestra una distribución logarítmica para ambos índices de conservación en las regiones que presentan SNPs validados, encontrándose alrededor del 5% de la distribución.

Luego se procedió a determinar las cruzas entre los conjuntos de variantes obtenidas y las variantes validadas.

**Tabla VII. Intersecciones entre los datos validados y los conjuntos de variantes obtenidas en cada paso caracterizado por su metodología correspondiente.**

| | **SNPs** | **MSV-3s** | **MSV-5s** |
|---|---|---|---|
| *MIVs (métodos clásicos)* | 3,49 % | 15,12 % | 26,87 % |
| *∂IVs* | 3,28 % | 14,27 % | 26,87 % |
| ***SNPs** Conservados (SNP-SACK)* | 1,34 % | 1,14 % | 0 % |

En la Tabla anterior se observa, como es de esperar que los métodos clásicos se inclinen a encontrar MSV-5s con una mayor probabilidad que MSV-3s y ambos a su vez con una mayor probabilidad que un SNP. Esto considerando que los MSVs pueden ser encontrados entre lecturas de un mismo individuo o entre lecturas de múltiples individuos de regiones genómicas distintas pero parálogas entre sí.

Los porcentajes de variantes validadas presentes en el conjunto de MIVs son similares a las variantes validadas presentes en el conjunto de IVs determinadas descartando en base a la densidad de IVs, tanto para SNPs como MSVs. Luego de aplicado SNP-SACK se observa una inversión en estos porcentajes.

Esta validación nos permite verificar la potencia de nuestro método predictivo en la discriminación de variantes conservadas con respecto a metodologías clásicas de búsqueda de variantes.

# 4 DISCUSIÓN

## *4.1 Pre-procesamiento del ensamble.*

La correlación lineal entre la longitud de RE y longitud de contigs observada en laFigura 15demuestra una distribución relativamente homogénea de RE a lo largo del ensamble utilizado, lo que indica que este parámetro no sería útil para generar discriminaciones a lo largo del ensamble(Belle et al., 2005). La abundancia de elementos repetitivos (59% de nuestra referencia), limita sustancialmente la búsqueda eficiente de marcadores en *Salmo salar*(Treangen and Salzberg, 2011), sin embargo la distribución homogénea de estos elementosmejora la posibilidad de obtención de un panel de marcadores con el nivel de densidad homogénea requerida para un estudio de selección genómica (1 c/ 60 Kb)(Hayes, 2007).

El ensamblepúblico utilizado, a pesar de tener solo un 89 % de contigs mayores a 1 Kb(Davidson et al., 2010), estos tienen una representacióndel 67%sobre el largo del genoma total estimado para *Salmo salar* (3.2Gb), traduciéndose en una disminución de tan solo un 15 % en la representatividad con respecto al ensamble total (2,4 Mb).Esto genera una nueva limitante en el número de marcadores posibles de obtener.

## *4.2 Obtención de marcadores candidatos.*

En la Tabla VIpodemos observar un menor porcentaje de mapeos de las lecturas pertenecientes a los 16 individuos chilenos, respecto a las lecturas de individuos noruegos (62% vs 72%), a pesar que estos últimos presentanuna cobertura 176 veces menor. Tomando en consideración que Sally es un individuo de una

población noruega es posible plantear la hipótesis de que las diferencias en mapeo se deban a las variaciones entre las cepas presentes en ambas poblaciones(Lubieniecki et al., 2010); no obstante cabe considerar que estos resultados son sensibles a la tecnología de secuenciación utilizada la que difiere entre ambas muestras(Metzker, 2010).

Del total de secuencias de múltiples individuos,sólo un 48,42% fueron mapeadas a una región única del ensamble, indicando que un gran porcentaje podría encontrarseen regiones que presentan parálogos con alta identidad(Moghadam et al., 2011) y elementos repetitivos que componen cerca del 60 % de nuestro genoma de referencia (de Boer et al., 2007; Kido et al., 1994).

La densidad de candidatos MIVs inicial obtenida (1 c/ 92 Kb), es insuficiente para generar un conjunto de marcadores para selección genómica(Hayes, 2007), haciendo necesario un aumento en la coberturautilizada para la búsqueda de MIVs (12,48X de cobertura). La cantidad de marcadores necesarios para selección genómica se estiman alrededor de 1 c/ 60 Kb (Hayes, 2007). Lo que hace necesario un conjuntomínimo de alrededor de 60 K marcadores equidistantes para la generación de un SNP-Chip con la densidad apropiada para la selección eficiente de caracteres de importancia económica mediante selección genómica. El conjunto de 30 K SNPs obtenido puede aportar con un primer acercamiento compuesto por ~ 14 K SNPs equidistantes que pueden ser seleccionados luego de un primer genotipado poblacional.

Los candidatos IVs obtenidos representan aproximadamente un 95 % de los PSVs/MSVs estimados por Hayes presentes en el genoma del salmón (~5 M)(Hayes et al., 2007). Esto tomando en consideración que solo se utilizó un ensamble con un 67% de representación del genoma de los salmónidos;lo anterior

se puede explicar por la presencia de otros tipos de variantes dentro del conjunto de IVs, generadas por errores de ensamble entre las secuencias de Sally por las dificultades de ensamble ya mencionadas (Davidson et al., 2010). Esto aún no es suficiente para explicar la alta presencia de IVs a pesar de la exclusión del 59% de la referencia debido a RE, lo que nos lleva a pensar que existiría una variabilidad mayor entre algunas de estas secuencias parálogas que la proyectada por Hayes(Hayes, 2007).

## *4.3 Predicción de estructuras génicas.*

Las estructuras génicas predichas contabilizan para un 80% de los genes estimados por UniGene en salmón del Atlántico(Pontius et al., 2003).Sin embargo se cuenta solo con un 46% de los UTR totales que generan que soloun 33 % de los genes predichos estén completos.La falta de secuencias no codificantes puede influir en el cálculo de los indicadores de conservación que toman en cuenta estas zonas (i.e. $\delta Var_{EXON}/\delta Var_{UTR}$). No obstante estas estructuras génicas nos permiten tener una estadística representativa de los índices de conservación a lo largo de las secuencias codificantes asociadas a los genes del salmón.

## *4.4 Caracterización de los polimorfismos.*

El archivo resumen de la caracterización de MIVs por SnpEff, mostró zonas con claros indicios de duplicación, como una aumentada densidad de marcadores(Hayes et al., 2007) (Figura 16). No obstante un descarte de variantes basado solo en la densidad de marcadores, produciría un enmascaramiento de grandes regiones del genoma, haciendo el conjunto final de marcadores obtenido por una metodología de esta naturaleza inviable para un estudio de selección genómica, el cual requiere de una distribución relativamente homogénea de marcadores a lo largo del genoma(Hayes, 2007). Por otro lado, nuestro método

descarta selectivamente regiones acotadas del genoma en base a su nivel de conservación, lo que permite generar un enmascaramiento más acotado y selectivo (Figura 19), necesario para la generación de un Chip para selección genómica.

## *4.5 Clusterización.*

La correlación lineal observada en la Figura 17 ($R^2$ = 0,97), nos indica que esta sería un buen predictor del rango de sets difusos posibles para los intervalos de ventanas móviles, a partir de los dos indicadores de conservación aplicables a estos (ver Tabla V).

Luego de la clusterización cerca del 7 % de los marcadores iniciales fueron descartados.La proporción $(MIVs - (MIVs \cap IVs)) : IVs$muestra un evidente aumento (34,41%) luego de la aplicación de nuestro método de agrupación basado en índices de conservación (SNP-SACK), demostrando la especificidad del método. Cabe recordar que los IVs además de PSVs también incluyen marcadores válidos MSVs, sin embargo en teoría deberían contener una mayor proporción de PSVs que el conjunto de MIVs, lo que se ve representado en el aumento de la proporción nombrada.

Se observa que de los marcadores finales cerca de un 73% se encontrarían en regiones intergénicas, lo que es de esperar ya que estas componen sobre un 99% del genoma del salmón del Atlántico.

El conjunto final de marcadores obtenido marca para una quinta parte del genoma del salmón del Atlántico, lo cual es suficiente como para realizar mapeo de QTLs y otras técnicas de selección(Hayes, 2007). Sin embargo se encuentra muy por debajo de lo requerido por selección genómica, haciendo necesario el

secuenciamiento de mayor número de lecturas genómicas de múltiples individuos, con el objeto de aumentar el número de MIVs inicial.

## *4.6 Validación in silico.*

En la validación *in silico* sobre un 93% de los True-SNPs validados cayeron dentro de zonas "clean"(conservadas).Evidenciando que SNP-SACK, a pesar de descartar zonas enriquecidas en PSVs, es capaz de conservar una gran parte de los MSVs presentes en la referencia. Lo anterior debido a que el criterio de clusterización se basa en nivel de conservación de secuencia y no en el descarte de la región duplicada completa.

Además, como es posible apreciar en la Tabla VII, SNP-SACK a pesar de tener porcentajes cercanos en el descarte de variantes validadas (Figura 17) tanto para SNPs como MSVs, descarta de manera selectiva MSVs que se encuentran más representados en los conjuntos de variantes determinadas por métodos tradicionales de búsqueda de variantes. Una explicación plausible para esto, es que aquellas variantes de menor conservación tendrán mayor representación en un conjunto de variantes determinadas al azar, implicando una alta efectividad de nuestro método de selección.

Estos resultados indican que el algoritmo propuesto y los parámetros utilizados son eficientespara la búsqueda de variantesútiles en la generación de un microarreglo de marcadores para selección genómica.

# 5 CONCLUSIONES

## *5.1 Proyecciones.*

Los principios tras el presente pipelinebioinformático de caracterización y búsqueda de marcadores (SNP-SACK), puede ser potencialmente aplicada a múltiples otros organismos con estados similares de caracterización genómica al estado actual del genoma del salmón del Atlántico.

La base de datos resultante contiene información acerca de la naturaleza de los polimorfismos descartados como PSVs, lo que permitie su potencial aplicación en otros estudios avanzados y Facilita la búsqueda sitio-específica de nuevos True-SNPs a través de la exploración de regiones del genoma caracterizadas como altamente conservadas por nuestro algoritmo. Además permite el análisis de las relaciones evolutivas de los genes encontrados en regiones con secuencias parálogas.

Un aumento en la disponibilidad de datos de marcadores probados permitiría potencialmente la generación de un conjunto de llegada. En ese escenario se recomendaría el uso de un modelo predictivo supervisado capaz de integrar criterios de experto para conjuntos difusos como Redes Neuronales Difusas (RND) (Buckley and Hayashi, 1994; Lee and Lee, 1975).

La disponibilidad de la colección de SNPs del genoma completo abriría nuevas posibilidades para explorar relaciones genéticas entre individuos y nos entrega las herramientas necesarias para analizar los componentes heredables y los genes determinantes de rasgos de importancia económica (QTLs), haciendo posible

realizar análisis comparativos generales de genoma completo entre individuos, así como análisis comparativos locales específicos entre regiones cromosomales. Estas metodologías pueden ser usadas para el mejoramiento comercial del salmón del Atlántico, su resistencia a patógenos, rastreo de líneas de parentesco y otras múltiples investigaciones relativas a la genómica comparativa.

Adicionalmente, a partir del mapa génico disponible (Quinn et al., 2008) y el conjunto de marcadores clasificados y mapeados para casa cada grupo cromosómico (Lien et al., 2011), sería posible inferir la organización en cromosomas de nuestro ensamble y los marcadores obtenidos.

La empresa Aquainnovo S.A. a través del financiamiento de la Corporación de Fomento de la Producción (CORFO) y del Programa Fondo de Investigación Avanzado en Áreas Prioritarias (FONDAP), pretende obtener secuencias genómicas de nuevos individuos ymediante el usode el algoritmo presentado en ésta tesis (SNP-SACK) generar de un SNP-Chip comercializable y con la densidad necesaria para tener una alta efectividad para la selección genómica del salmón del Atlántico.

## *5.2 Conclusión final.*

Un conjunto de marcadores génicos de alta densidad es esencial para el desarrollo de un SNPchip eficiente para la selección genómica.

El Kit bioinformático de herramientas desarrollado "SNP-SACK",a través de un acercamiento innovador, permite una clasificación eficiente de SNPs *a priori*acorde con los niveles de conservación de la región del genoma en la que estén ubicados. Los datos resultantes de SNP-SACKfacilitan la selección de marcadores altamente conservados para el diseño eficiente de un SNPchip. La especificidad y sensibilidad de SNP-SACK es altamente dependiente de los parámetros utilizados y estos

pueden ser fácilmente regulados y ajustados alas necesidades del estudio a realizar.

La densidad de marcadores obtenida en el presente análisis se encuentra por debajo del ideal necesario para la generación de un SNPchip para selección genómica, por lo que sería necesaria la secuenciación de un mayor número de lecturas genómicas de múltiples individuos para aumentar la densidad de marcadores putativos iniciales.

A partir del procedimiento anterior, y los desarrollos metodológicos propios que se realizaron en función de los resultados obtenidos en cada etapa, se obtuvo un conjunto de SNPs/DIPs eficiente para su utilización en estudios de análisis de genoma completo y otros análisis de genómica comparativa en salmones,validados *in sillico* contra un subconjunto de True-SNPs.

Nuestra metodología representa un acercamiento único que prueba ser efectivo para la búsqueda, caracterización y filtrado de SNPs/DIPs, permitiendo su diferenciación de potenciales PSVs*a priori* para su aplicación en selección genómica.

# 6 BIBLIOGRAFÍA

© FAO - Fisheries and Aquaculture Information and Statistics Service (2012). FIGIS - Time-series query on: Aquaculture.

Abdi, H., and Williams, L.J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics *2*, 433–459.

Adzhubei, A., Vlasova, A., Hagen-Larsen, H., Ruden, T., Laerdahl, J., and HU00F8yheim, 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. (1984). Tetraploidy and the evolution of salmonid fishes. Evolutionary Genetics of Fishes.

Allendorf, F.W., and Danzmann, R.G. (1997). Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. Genetics *145*, 1083.

Allendorf, F.W., and Utter, F.M. (1973). Gene duplication within the family Salmonidae: Disomic inheritance of two loci reported to be tetrasomic in rainbow trout. Genetics *74*, 647–654.

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

Andersson, L. (2001). Genetic dissection of phenotypic diversity in farm animals. Nature Reviews Genetics *2*, 130–138.

Andersson, L., and Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews Genetics *5*, 202–212.

Andreassen, R., Lunner, S., and Høyheim, B. (2009). Characterization of full-length sequenced cDNA inserts (FLIcs) from Atlantic salmon (Salmo salar). BMC Genomics *10*, 502.

Andreassen, R., Lunner, S., and Høyheim, B. (2010). Targeted SNP discovery in Atlantic salmon (Salmo salar) genes using a 3'UTR-primed SNP detection approach. BMC Genomics *11*, 706.

Bagnato, A., and Rosati, A. (2012). From the Editors—Animal selection: The genomics revolution. Animal Frontiers *2*, 1–2.

Bailey, G.S., Poulter, R., and Stockwell, P.A. (1978). Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. Proceedings of the National Academy of Sciences *75*, 5575.

Baker, M. (2012). De novo genome assembly: what every biologist should know. Nature Methods *9*, 333–337.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. (2002). ARACHNE: a whole-genome shotgun assembler. Genome Research *12*, 177–189.

Belle, E.M.S., Webster, M.T., and Eyre-Walker, A. (2005). Why are young and old repetitive elements distributed differently in the human genome? J. Mol. Evol. *60*, 290–296.

Bennewitz, J., Reinsch, N., Szyda, J., Reinhardt, F., Kuhn, C., Schwerin, M., Erhardt, G., Weimann, C., and Kalm, E. (2003). Marker assisted selection in German Holstein dairy cattle breeding: outline of the program and marker assisted breeding value estimation. Book of Abstr. 54th Annu. Mtg. Eur. Assoc. Anim. Prod. Y. van Der Honing, Ed. Wageningen Academic Publishers, Wageningen, The Netherlands 5.

Benson, D.A., Boguski, M.S., Lipman, D.J., and Ostell, J. (1997). GenBank. Nucl. Acids Res. *25*, 1–6.

Bernardi, G. (1989). The Isochore Organization of the Human Genome. Annual Review of Genetics *23*, 637–659.

Blanca, J.M., Pascual, L., Ziarsolo, P., Nuez, F., and Cañizares, J. (2011). ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using Next Generation Sequence. BMC Genomics *12*, 285.

De Boer, J., Yazawa, R., Davidson, W., and Koop, B. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics *8*, 422.

Boguski, M.S., Lowe, T.M.J., and Tolstoshev, C.M. (1993). dbEST |[mdash]| database for |[ldquo]|expressed sequence tags|[rdquo]|. Nature Genetics *4*, 332–333.

Boichard, D., Fritz, S., Rossignol, M.N., Boscher, M.Y., Malafosse, A., and Colleau, J.J. (2002). Implementation of marker-assisted selection in French dairy cattle. In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August, 2002. Session 22., (Institut National de la Recherche Agronomique (INRA)), pp. 1–4.

Brownstein, M.J., Carpten, J.D., and Smith, J.R. (1996). Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. Biotechniques *20*, 1004–1006.

Bucciarelli, G., Bernardi, G., and Bernardi, G. (2002). An ultracentrifugation analysis of two hundred fish genomes. Gene *295*, 153–162.

Buckley, J.J., and Hayashi, Y. (1994). Fuzzy neural networks: A survey. Fuzzy Sets and Systems *66*, 1–13.

Burrows, M., and Wheeler, D.J. (1994). A Block-Sorting Lossless Data Compression Algorithm. Digital Systems Research Center. RR-124.

Byvatov, E., and Schneider, G. (2003). Support vector machine applications in bioinformatics. Applied Bioinformatics *2*, 67.

Chapman, B., and Chang, J. (2000). Biopython: Python tools for computational biology. ACM SIGBIO Newsletter *20*, 15–19.

Chicurel, M. (2002). Bioinformatics: Bringing it all together technology feature. Nature *419*, 751–757.

Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M.,

and Lu, X. (2012a). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Frontiers in Genetics *3*.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Ruden, D.M., and Lu, X. (2012b). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 0–1.

Closter, A.M., Elferink, M.G., As, P. van, Crooijmans, R., Groenen, M.A.M., and Bovenhuis, H. (2010). Genome-wide association analysis identifies loci that influence ascites in broilers. In 9th World Congress on Genetics Applied to Livestock Production, Leipzig, p. 4.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucl. Acids Res. *38*, 1767–1771.

Colleau, J.J., Fritz, S., Guillaume, F., Baur, A., Dupassieux, D., Boscher, M.Y., Journaux, L., Eggen, A., and Boichard, D. (2009). Simulating the potential of genomic selection in dairy cattle breeding. Rencontres Recherches Ruminants *16*, 419.

Collins, D.W., and Jukes, T.H. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics *20*, 386–396.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The Variant Call Format and VCFtools. Bioinformatics *27*, 2156–2158.

Darvasi, A., and Soller, M. (1997). A Simple Method to Calculate Resolving Power and Confidence Interval of QTL Map Location. Behavior Genetics *27*, 125–132.

Davey, G.C., Caplice, N.C., Martin, S.A., and Powell, R. (2001a). A survey of genes in the Atlantic salmon (Salmo salar) as identified by expressed sequence tags. Gene *263*, 121.

Davey, G.C., Caplice, N.C., Martin, S.A., and Powell, R. (2001b). A survey of genes in the Atlantic salmon (Salmo salar) as identified by expressed sequence tags. Gene *263*, 121–130.

Davidson, W.S., Huang, T.-K., Fujiki, K., von Schalburg, K.R., and Koop, B.F. (2009). The sex determining loci and sex chromosomes in the family salmonidae. Sex Dev *3*, 78–87.

Davidson, W.S., Koop, B.F., Jones, S.J.M., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S., Omholt, S.W., and others (2010). Sequencing the genome of the Atlantic salmon (Salmo salar). Genome Biol *11*, 403.

Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Commun. ACM *51*, 107–113.

Dehal, P., and Boore, J.L. (2005). Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol *3*, e314.

Dekkers, J.C.M. (2004). Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. Journal of Animal Science *82*, E313–E328.

Dekkers, J.C.M., and Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics *3*, 22–32.

Delgado, C.L., Wada, N., Rosegrant, M.W., Meijer, S., and Ahmed, M. (2003). Fish to 2020. World Fish Centre and International Food Policy Research Institute, Penang, Malaysia.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dominik, S., Henshall, J.M., Kube, P.D., King, H., Lien, S., Kent, M.P., and Elliott, N.G. (2010). Evaluation of an Atlantic salmon SNP chip as a genomic tool for the application in a Tasmanian Atlantic salmon (Salmo salar) breeding population. Aquaculture *308, Supplement 1*, S56–S61.

Earl, D., Bradnam, K., John, J.S., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., et al. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Research *21*, 2224–2241.

Ewing, B., Green, P., and others (2000). Analysis of expressed sequence tags indicates 35,000 human genes. Nature Genetics *25*, 232–234.

FAO (2011). Mapping supply and demand for animal-source foods to 2030, by T.P. Robinson & F. Pozzi. Animal Production and Health Working Paper. No. 2. Rome. ISSN 2221-8793 16–26.

FAO (2012). Aquaculture Department.(2012) The state of world fisheries and aquaculture. FAO Fisheries Technical Paper. Rome, Italy. ISSN 1020–5489.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2010). Ensembl 2011. Nucleic Acids Research *39*, D800–D806.

Fredman, D., White, S.J., Potter, S., Eichler, E.E., Dunnen, J.T.D., and Brookes, A.J. (2004). Complex SNP-related sequence variation in segmental genome duplications. Nature Genetics *36*, 861–866.

Gath, I., and Geva, A.B. (1989). Unsupervised optimal fuzzy clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions On *11*, 773–780.

Di Genova, A., Aravena, A., Zapata, L., Gonzalez, M., Maass, A., and Iturra, P. (2011). SalmonDB: a bioinformatics resource for Salmo salar and Oncorhynchus mykiss. Database *2011*, bar050–bar050.

Di Génova, A., Aravena, A., Zapata, L., González, M., Maass, A., and Iturra, P. (2011). SalmonDB: a bioinformatics resource for Salmo salar and Oncorhynchus mykiss. Database (Oxford) *2011*.

Gidskehaug, L., Kent, M., Hayes, B.J., and Lien, S. (2011). Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array.

Bioinformatics *27*, 303–310.

Ginot, F., Bordelais, I., Nguyen, S., and Gyapay, G. (1996). Correction of some genotyping errors in automated fluorescent microsatellite analysis by enzymatic removal of one base overhangs. Nucleic Acids Res. *24*, 540–541.

Golding, B., Morton, D., and Haerty, W. (2000). Elementary Sequence Analysis. Multiple Sequence Alignments.

Gosling, J., and McGilton, H. (1995). The Java language environment (Sun Microsystems Computer Company).

Guo, Y., and Jamison, D.C. (2005). The distribution of SNPs in human gene regulatory regions. BMC Genomics *6*, 140.

Gutiérrez, A.P., Lubieniecki, K.P., Davidson, E.A., Lien, S., Kent, M.P., Fukui, S., Withler, R.E., Swift, B., and Davidson, W.S. (2012). Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (Salmo salar) using a 6.5 K SNP array. Aquaculture *358–359*, 61–70.

Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E.E., and Sahinalp, S.C. (2010). mrsFAST: a cache-oblivious algorithm for short-read mapping. Nature Methods *7*, 576–577.

Hagen-Larsen, H., Laerdahl, J.K., Panitz, F., Adzhubei, A., and Høyheim, B. (2005). An EST-based approach for identifying genes expressed in the intestine and gills of pre-smolt Atlantic salmon (Salmo salar). BMC Genomics *6*, 171.

Hahn, M., Wilhelm, J., and Pingoud, A. (2001). Influence of fluorophor dye labels on the migration behavior of polymerase chain reaction--amplified short tandem repeats during denaturing capillary electrophoresis. Electrophoresis *22*, 2691–2700.

Hardie, D.C., and Hebert, P.D.. (2003). The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. Genome *46*, 683–706.

Hayes, B. (2007). QTL mapping, MAS, and genomic selection. A Short-course Organized by Animal Breeding & Genetics Department of Animal Science, Iowa State University.

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. Genome *53*, 876–883.

Hayes, B., Goddard, M.E., and others (2001). The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution *33*, 209–230.

Hayes, B., Laerdahl, J., Lien, S., Moen, T., Berg, P., Hindar, K., Davidson, W., and Koop, B. (2007). An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (Salmo salar) expressed sequences. Aquaculture *265*, 82–90.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., and Goddard, M.E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science *92*, 433–443.

Huang, S., Xu, X., and Pan, S.K. (2011). Genome sequence and analysis of

the tuber crop potato. Nature *475*, U189–U194.

Huang, X., Wang, J., Aluru, S., Yang, S.-P., and Hillier, L. (2003). PCAP: a whole-genome assembly program. Genome Research *13*, 2164–2170.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucl. Acids Res. *30*, 38–41.

Hume, D.A., Whitelaw, C.B.A., and Archibald, A.L. (2011). The future of animal production: improving productivity and sustainability. The Journal of Agricultural Science *149*, 9–16.

Hyten, D.L., Cannon, S.B., Song, Q., Weeks, N., Fickus, E.W., Shoemaker, R.C., Specht, J.E., Farmer, A.D., May, G.D., and Cregan, P.B. (2010a). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. Bmc Genomics *11*, 38.

Hyten, D.L., Cannon, S.B., Song, Q., Weeks, N., Fickus, E.W., Shoemaker, R.C., Specht, J.E., Farmer, A.D., May, G.D., and Cregan, P.B. (2010b). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics *11*, 38.

Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 299–314.

Johnson, K., Wright, J., and May, B. (1987). Linkage relationships reflecting

Genomics *9*, 545.

ancestral tetraploidy in salmonid fish. Genetics *116*, 579.

Jr, W.J., K, J., A, H., and B, M. (1983). Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. Isozymes *10*, 239.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research *110*, 462–467.

Kasahara, M. (2007). The 2R hypothesis: an update. Current Opinion in Immunology *19*, 547–552.

Kido, Y., Himberg, M., Takasaki, N., and Okada, N. (1994). Amplification of Distinct Subfamilies of Short Interspersed Elements During Evolution of the Salmonidae. Journal of Molecular Biology *241*, 633–644.

Kim, H., Schmidt, C.J., Decker, K.S., and Emara, M.G. (2003). A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. Animal Genetics *34*, 249–254.

Klir, G.J., and Yuan, B. (1995). Fuzzy sets and fuzzy logic (Prentice Hall New Jersey).

Koop, B., Schalburg, K. von, Leong, J., Walker, N., Lieph, R., Cooper, G., Robb, A., Beetz-Sargent, M., Holt, R., Moore, R., et al. (2008). A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. BMC Genomics *9*, 545.

Lam, H.Y.K., Pan, C., Clark, M.J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M.B., Kidd, J.M., Bustamante, C.D., et al. (2012). Detecting and annotating genetic variations using the HugeSeq pipeline. Nature Biotechnology *30*, 226–229.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359.

Laurenti, G. (2008). Fish and fishery products: world apparent consumption statistics based on food balance sheets (FAO).

Lee, S.C., and Lee, E.T. (1975). Fuzzy neural networks. Mathematical Biosciences *23*, 151–177.

Leong, J., Jantzen, S., Schalburg, K. von, Cooper, G., Messmer, A., Liao, N., Munro, S., Moore, R., Holt, R., Jones, S., et al. (2010). Salmo salar and Esox lucius full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. BMC Genomics *11*, 279.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The Diploid Genome Sequence of an Individual Human. PLoS Biol *5*, e254.

Li, H. (2011). Improving SNP Discovery by Base Alignment Quality. Bioinformatics *27*, 1157–1158.

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. Bioinformatics *25*, 1754–1760.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics *26*, 589 –595.

Li, S., and Chou, H.H. (2004). LUCY2: an interactive DNA sequence quality trimming and vector removal tool. Bioinformatics *20*, 2865–2866.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J., Boroevich, K., Koop, B., and Davidson, W. (2011). Comparative Genomics Identifies Candidate Genes for Infectious Salmon Anemia (ISA) Resistance in Atlantic Salmon (<i>Salmo salar</i>). Marine Biotechnology *13*, 232–241.

Li, W.-H., Gojobori, T., and Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. , Published Online: 16 July 1981; | Doi:10.1038/292237a0 *292*, 237–239.

Lien, S., Gidskehaug, L., Moen, T., Hayes, B., Berg, P., Davidson, W., Omholt, S., and Kent, M. (2011). A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics *12*, 615.

Lipman, D.J., Pearson, W.R., and others (1985). Rapid and sensitive protein similarity searches. Science *227*, 1435–1441.

Lubieniecki, K.P., Jones, S.L., Davidson, E.A., Park, J., Koop, B.F., Walker,

S., and Davidson, W.S. (2010). Comparative genomic analysis of Atlantic salmon, Salmo salar, from Europe and North America. BMC Genetics *11*, 105.

Lynch, M., Walsh, B., and others (1998). Genetics and analysis of quantitative traits.

Mardis, E.R., and others (2008). The impact of next-generation sequencing technology on genetics. Trends in Genetics *24*, 133.

Martin, S.A., Caplice, N.C., Davey, G.C., and Powell, R. (2002). EST-based identification of genes expressed in the liver of adult Atlantic salmon (Salmo salar). Biochem. Biophys. Res. Commun. *293*, 578–585.

Martínez-Arias, R., Calafell, F., Mateu, E., Comas, D., Andrés, A., and Bertranpetit, J. (2001). Sequence variability of a human pseudogene. Genome Res. *11*, 1071–1085.

May, B.P. (1980). The salmonid genome: evolutionary restructuring following a tetraploid event. Pennsylvania State University.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research *20*, 1297–1303.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat Rev Genet *11*, 31–46.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001a). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics *157*, 1819–1829.

Meuwissen, T.M.H., Goddard, M.E., and others (2001b). Prediction of identity by descent probabilities from marker-haplotypes. Genetics Selection Evolution *33*, 605–634.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive Assembly of Pyrosequencing Reads with Mates. Bioinformatics *24*, 2818–2824.

Moen, T., Hayes, B., Baranski, M., Berg, P., KjU00F8glum, S., Koop, B., Davidson, W., Omholt, S., and Lien, S. (2008). A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. BMC Genomics *9*, 223.

Moghadam, H.K., Ferguson, M.M., and Danzmann, R.G. (2011). Whole genome duplication: challenges and considerations associated with sequence orthology assignment in Salmoninae. Journal of Fish Biology *79*, 561–574.

Myers, E.W. (2000). A Whole-Genome Assembly of Drosophila. Science *287*, 2196–2204.

Myers, R.A., Worm, B., and others (2003). Rapid worldwide depletion of predatory fish communities. Nature *423*, 280–283.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot,

B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Ng, S.H.S., Artieri, C.G., Bosdet, I.E., Chiu, R., Danzmann, R.G., Davidson, W.S., Ferguson, M.M., Fjell, C.D., Hoyheim, B., Jones, S.J.M., et al. (2005). A physical map of the genome of Atlantic salmon, Salmo salar. Genomics *86*, 396–404.

Norman, J.D., Robinson, M., Glebe, B., Ferguson, M.M., and Danzmann, R.G. (2012). Genomic arrangement of salinity tolerance QTLs in salmonids: A comparative analysis of Atlantic salmon (Salmo salar) with Arctic charr (Salvelinus alpinus) and rainbow trout (. BMC Genomics *13*, 420.

Nystøyl, R. (2011). Salmon Production Review.

Ohno, S., and others (1970). Evolution by gene duplication. (London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.).

Panetta, L.E. (2003). America's living oceans: charting a course for sea change: a report to the nation: recommendations for a new ocean policy (Pew Oceans Commission).

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. *85*, 2444–2448.

Pevzner, P.A., and Tang, H. (2001). Fragment assembly with double-barreled data. Bioinformatics *17*, S225–S233.

Phillips, R., and Ráb, P. (2001). Chromosome evolution in the Salmonidae (Pisces): an update. Biol Rev Camb Philos Soc *76*, 1–25.

Plastow, G., Sasaki, S., Yu, T., Deeb, N., Prall, G., Siggens, K., and Wilson, E. (2003). Practical application of DNA markers for genetic improvement. Proc. 28th Annu. Mtg. Natl. Swine Improve. Fed., Iowa State Univ., Ames 151–154.

Pontius, J.U., Wagner, L., and Schuler, G.D. (2003). 21. UniGene: A Unified View of the Transcriptome. The NCBI Handbook. Bethesda, MD: National Library of Medicine (US), NCBI.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research *35*, D61–D65.

Quinn, N.L., Levenkova, N., Chow, W., Bouffard, P., Boroevich, K.A., Knight, J.R., Jarvie, T.P., Lubieniecki, K.P., Desany, B.A., Koop, B.F., et al. (2008). Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. BMC Genomics *9*, 404.

Reis-Filho, J.S., and others (2009). Next-generation sequencing. Breast Cancer Res *11*, S12.

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., et al. (2004). Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics.

Genome Res. *14*, 478–490.

Roberts, F.L. (1970). Atlantic Salmon (Salmo salar) Chromosomes and Speciation. Transactions of the American Fisheries Society *99*, 105–111.

Rohrer, G.A., Freking, B.A., and Nonneman, D. (2007). Single nucleotide polymorphisms for pig identification and parentage exclusion. Animal Genetics *38*, 253–258.

Rosegrant, M.W., Cai, X., and Cline, S.A. (2002). World water and food to 2025: Dealing with scarcity (International Food Policy Research Inst).

Ross, T.J. (2010). Front Matter, Front Matter. In Fuzzy Logic with Engineering Applications, Third Edition, Fuzzy Logic with Engineering Applications, (John Wiley & Sons, Ltd, John Wiley & Sons, Ltd), pp. i, i–xxi, xxi.

Rothschild, M.F., Larson, R.G., Jacobson, C., and Pearson, P. (1991). PvuII polymorphisms at the porcine oestrogen receptor locus (ESR). Anim. Genet. *22*, 448.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Research *22*, 557–567.

Sánchez, C., Smith, T., Wiedmann, R., Vallejo, R., Salem, M., Yao, J., and Rexroad, C. (2009). Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC

Genomics *10*, 559.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences *74*, 5463–5467.

Sanner, M.F. (1999). Python: a programming language for software integration and development. J Mol Graph Model *17*, 57–61.

Sargolzaei, M., Schenkel, F.S., and VanRaden, P.M. (2009). gebv: Genomic breeding value estimator for livestock (Technical report to the Dairy Cattle Breeding and Genetics Committee).

Sato, Y., and Nishida, M. (2010). Teleost fish with specific genome duplication as unique models of vertebrate evolution. Environmental Biology of Fishes *88*, 169–188.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nature *200*, 16–18.

Schwab, M., Karrenbach, N., and Claerbout, J. (2000). Making scientific computations reproducible. Computing in Science Engineering *2*, 61 –67.

Secko, D., Ilves, K., and Burgess, M. (2007). Report (Y1): cGRASP–GE3LS Project, Issues and Priorities.

Seguí-Simarro, J.M., and Nuez, F. (2008). Pathways to doubled haploidy: chromosome doubling during androgenesis. Cytogenetic and Genome Research *120*, 358–369.

Serpell, J. (1996). In the Company of Animals: A Study of Human-Animal Relationships (Cambridge University Press).

Shen, J.-C., Rideout, W.M., and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucl. Acids Res. *22*, 972–976.

Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: The NCBI Database of Genetic Variation. Nucl. Acids Res. *29*, 308–311.

Shrimpton, A., and Robertson, A. (1988). The isolation of polygenic factors controlling bristle score in Drosophila melanogaster. II. Distribution of third chromosome bristle effects within chromosome sections. Genetics *118*, 445–459.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. Genome Research *19*, 1117–1123.

Siva, N. (2008). 1000 Genomes project. Nature Biotechnology *26*, 256–256.

Smit, A., Hubley, R., and Green, P. (2004). RepeatMasker Open-3.0.

Smith, C.T., Elfstrom, C.M., Seeb, L.W., and Seeb, J.E. (2005). Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Molecular Ecology *14*, 4193–4203.

Smith, E.J., Shi, L., Drummond, P., Rodriguez, L., Hamilton, R., Ramlal, S.,

Smith, G., Pierce, K., and Foster, J. (2001). Expressed sequence tags for the chicken genome from a normalized 10-day-old White Leghorn whole embryo cDNA library: 1. DNA sequence characterization and linkage analysis. J. Hered. *92*, 1–8.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., et al. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Res. *12*, 1611–1618.

Stallman, R.M., and McGrath, R. (1988). GNU make (Free Software Foundation).

Stein, L. (1996). How Perl saved the human genome project. Dr Dobb's Journal (July 2001).

Steinke, D., Salzburger, W., and Meyer, A. (2006). Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. J. Mol. Evol. *62*, 772–784.

Stern, C. (1943). The Hardy-Weinberg Law. Science *97*, 137–138.

Strachan, T., and Read, A.P. (2004). Human Molecular Genetics 3 (Garland Science).

Subasinghe, R.P., Bondad-Reantaso, M.G., and McGladdery, S.E. (2001). Aquaculture development, health and wealth.

Svärdson, G. (1945). Chromosome studies on Salmonidae. Ivar Hoeggströms.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. In Current Protocols in Bioinformatics, A.D. Baxevanis, G.A. Petsko, L.D. Stein, and G.D. Stormo, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.),.

Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods *5*, 247–252.

Thorsen, J., Zhu, B., Frengen, E., Osoegawa, K., de Jong, P., Koop, B., Davidson, W., and Høyheim, B. (2005). A highly redundant BAC library of Atlantic salmon (Salmo salar): an important tool for salmon projects. BMC Genomics *6*, 50.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics *13*, 36–46.

Tsoi, S.C.M., Ewart, K.V., Penny, S., Melville, K., Liebscher, R.S., Brown, L.L., and Douglas, S.E. (2004). Identification of immune-relevant genes from atlantic salmon using suppression subtractive hybridization. Mar. Biotechnol. *6*, 199–214.

Utter, F.M., Allendorf, F.W., and Hodgins, H.O. (1973). Genetic Variability and Relationships in Pacific Salmon and Related Trout Based on Protein Variations. Systematic Biology *22*, 257 –270.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A., and others (2002). A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution *34*, 275–306.

Wall, L., and others (1994). The Perl programming language (Prentice Hall Software Series).

Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. Science *280*, 1077–1082.

Weber, J.L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. (2002). Human Diallelic Insertion/Deletion Polymorphisms. The American Journal of Human Genetics *71*, 854–862.

Who, J., and Consultation, F.E. (2003). Diet, nutrition and the prevention of chronic diseases. WHO Technical Report Series *916*, 3. Global and regional food consumption patterns and trends.

Wiedmann, R.T., Smith, T.P.L., and Nonneman, D.J. (2008). SNP discovery in swine by reduced representation and high throughput pyrosequencing. Bmc Genetics *9*, 81.

Wives, L.K., and Loh, S. (1998). Hyperdictionary: a knowledge discovery tool

to help information retrieval. In String Processing and Information Retrieval: A South American Symposium, 1998. Proceedings, pp. 103–109.

Wondji, C., Hemingway, J., and Ranson, H. (2007). Identification and analysis of single nucleotide polymorphisms (SNPs) in the mosquito Anopheles funestus, malaria vector. BMC Genomics *8*, 5.

Wright, J.E., May, B., Stoneking, M., and Lee, G.M. (1980). Pseudolinkage of the Duplicate Loci for Supernatant Aspartate Aminotransferase in Brook Trout, Salvelinus Fontinalis. J Hered *71*, 223–228.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859.

Wyszynski, M., Gabbara, S., and Bhagwat, A.S. (1994). Cytosine deaminations catalyzed by DNA cytosine methyltransferases are unlikely to be the major cause of mutational hot spots at sites of cytosine methylation in Escherichia coli. Proceedings of the National Academy of Sciences *91*, 1574.

Yang, Z., and Yoder, A.D. (1999). Estimation of the Transition/Transversion Rate Bias and Species Sampling. Journal of Molecular Evolution *48*, 274–283.

Ye, J., Parra, E.J., Sosnoski, D.M., Hiester, K., Underhill, P.A., and Shriver, M.D. (2002). Melting Curve SNP (McSNP) Genotying: a Useful Approach for Diallelic Genotyping in Forensic Science. Journal of Forensic Sciences *47*,

593–600.

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

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. Journal of Computational Biology *7*, 203–214.

Zhao, Z., and Boerwinkle, E. (2002). Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome. Genome Res. *12*, 1679–1686.