Análisis automatizado de Dominios ancestrales asociados al sistema inmune en cordados basales dentro de un contexto evolutivo

Camilo Alejandro Cerón Noriega

Universidad Nacional de Colombia
Facultad de Ingeniería
Bogota D.C, Colombia
2017

Análisis automatizado de Dominios ancestrales asociados al sistema inmune en cordados basales dentro de un contexto evolutivo

Camilo Alejandro Cerón Noriega

Tesis presentada como requisito parcial para optar al título de:
Magister en Bioinformática

Directora:
Dr. rer. nat. Clara Isabel Bermudez Santana
Departamento de Biología, Facultad de Ciencias
Universidad Nacional de Colombia

Línea de Investigación:
Inmunología evolutiva y Bioinformática
Grupo de Investigación:
Rnómica Teórica y Computacional

Universidad Nacional de Colombia
Facultad de Ingeniería
Bogota D.C, Colombia
2017

Al silencio

Agradecimientos

- A la Profesora Clara Isabel Bermúdez por todo su apoyo incondicional durante este proceso.
- A su grupo de investigación Grupo Rnomica Teórica y Computacional en especial a Cristian Velandia, Andres Puerta y Oscar Escobar por su apoyo y paciencia durante la ejecución de este trabajo.
- Se agradece a Colciencias por el financiamiento parcial de mi trabajo de Maestría dentro del marco del proyecto financiado por Colciencias *"Estudio de la organización genómica de regiones adyacentes a dominios ancestrales del sistema inmune de cordados inferiores"*, liderado por el grupo RNómica Teórica y Computacional de la Universidad Nacional de Colombia en conjunto con pares de la Universidad de Leipzig (Alemania), la Universidad de São Paulo (Brasil) y el Instituto GiMaRIS en Holanda.
- Se agradece al Profesor Arjan Gittenberger del Instituto GiMARIS de Holanda por haber permitido el uso del genoma borrador de la especie *Didemnum vexillum* para lograr el objetivo 2.
- Al Profesor Peter F. Stadler y al Ingeniero Jens Steuck de la Universidad de Leipzig por su apoyo en el uso de infraestructura de computo robusto para el cálculo de algunas mediciones realizadas en este trabajo.
- Se agradece al DAAD por proporcionar el laboratorio de Biología Computacional de la Facultad de Ciencias de la Universidad Nacional de Colombia en donde fue posible el desarrollo del mayor conjunto de procesos usados.
- Expreso mi gratitud a COLCIENCIAS por haber permitido ser co-investigador asociado y permitir mi concentración en el desarrollo de todas la actividades de mi tesis.

Resumen

Esta tesis tiene como objetivo la construcción de un flujo automatizado de trabajo que integre diferentes procesos, la información de bases de datos y los modelos computacionales requeridos para identificar dominios asociados al Sistema Inmune (SI) presente en los tunicados que se conoce como un repertorio de genes principalmente asociado al Sistema Inmune Innato(SII). La importancia bioinformática de este trabajo se fundamenta en la necesidad de reconstruir un modelo de ganancia y perdida de dominios del SI en los tunicados bajo una aproximación de procesos automatizados aplicados sobre los genomas de algunas especies. Como grupo cercano a los vertebrados el screening promete revelar información de interés sobre la evolución del SI debido a que los tunicados se encuentran justo antes del bigbang inmunológico que es un proceso que se considera dio origen a la complejidad del Sistema Inmune Adaptativo. Por tanto desde el punto de vista de la rama de la bioinformática de anotación de genes, este trabajo propone una alternativa para la reconstrucción de regiones codificantes en especies no modelo que carecen de información de datos de expresión centrada en homología de dominios. La mayoria de las herramientas disponibles en anotación son altamente dependientes de la información transcriptómica o proteómica aunque existen métodos *ab initio* que se fudamentan en la búsqueda de señales propias de las secuencias de los genes procariotes y eucariotes. Por otro lado, en particular, la anotación de los genes del Sistema Inmune tampoco es sencilla ya que se debe atacar el problema computacional de identificar homología entre secuencias que pueden estar cargadas de ganancia, perdida y rearreglos de dominios. Es por esto que se propone en este trabajo esta estrategia que combina arquitecturas de dominios canónicos de genes del SII para una eficiente detección de dominios ultraconservados entre las especies de estudio. La estrategia se diseña con el fin identificar tractos de dominios en especies de tunicados que carecen de datos de transcriptómica o proteómica y por tanto se propone un modelo de identificación de posibles regiones genomicas putativas asociadas a codificar para dominios del SII en el genoma borrador de una especie carente de anotación y de datos de expresión como el tunicado *Didemnum vexillum* .

Finalmente, se implementa un modelo evolutivo de ganancia y perdida de dominios ultraconservados de genes putativos del SII. Dicha pipeline fue aplicada sobre la totalidad de los genomas de cinco especies de tunicados y de un grupo externo conformado por un cefalocordado y dos vertebrados.

Las características de los genomas evaluados durante esta tesis, en especial la de los tunicados, representaron retos computacionales importantes de tres tipos: primero genomas con peculiares historias evolutivas, segundo para algunas de estas especies los ensambles de los genomas se encuentran altamente fragmentados y como no son todos ellos organismos modelo no cuentan con información experimental amplia que permita entrenar y utilizar programas de anotación de genes ampliamente usados en Cordados como la pipeline de Ensembl y tercero existe complejidad en la arquitectura génica de los genes del SI ya que en ellos se presentan duplicaciones de dominios, rearreglos de los mismos y per-

didas. Estos problemas fueron resueltos en el Capítulo 1 mediante un análisis focalizado en el amplio repertorio de la arquitectura de genes existentes en dos bases de datos principales `InnateDB` y `Insect Innate Immunity Database` (IIID) usado para definir un sistema de dominios "Gold Standard" sobre las especies articuladas en el Ensembl usando BioMart para ser mapeados sobre las especies `Ciona intestinalis, Ciona savignyi, Petromyzon marinus, Latimeria chalumnae y Danio rerio` logrando así identificar el conjunto de dominios del SII de cordados inferiores. Posteriormente para las especies de tunicados `Oikopleura dioica y Botryllus schlosseri` y el `protocordado Branchiostoma floridaeo` que carecen de anotación de la pipeline del Ensembl se usaron las secuencias de sus proteínas reportadas, como blancos para la identificación de dominios canónicos asociados al SII previamente establecido. En el capitulo 2 se presenta la estrategia utilizada para identificar dominios en especies que carecen de evidencia experimental de expresión y anotación como el tunicado *D. vexillum* . Esta restricción en el número de dominios evaluados permitió de forma rápida, precisa y eficiente establecer conjuntos de dominios con arquitecturas proteicas similares a las reportadas en la literatura, siendo éstas el punto de partida para la búsqueda de relaciones de homología, principalmente de ortología y paralogía y de un modelo de ganancias y perdidas de dominios ultraconservados del SII descrito en el Capítulo 3.

Palabras clave: Dominios, Sistema Inmune, Sistema Inmune Innato, Tunicados, Anotación de Genes, Ganancia y Perdida.

Resumen

This survey is aimed at to build an automated workflow that integrates different processes, database information and computational models required to identify domains associated with the Immune System (IS) present in the tunicates that is known as a repertoire of genes mainly associated with the Innate Immune System (IIS). The bioinformatic importance of this work is based on the need of build a model of gain and loss of domains of the IS in tunicates following an approach which relies on an automated processes applied to the genomes of some species. As a group close to vertebrates, the screening on tunicates promises to reveal information of interest on the evolution of the IS because the tunicates are located just before the immunological big bang which is a process considered to have given rise to the complexity of the Adaptive Immune System. Therefore, from the point of view of the bioinformatics branch of gene annotation, this work proposes an alternative approach to reconstruct coding regions in non-model species that lack gene expression data centered on domain homology search since most of the tools available for gene annotation are highly dependent on the transcriptomic or proteomic information, although there are methods *ab initio* centered on the search for signals of the prokaryotic and eukaryotic genes sequences. On the other hand, in particular, the annotation of the genes of the Immune System is not a simple task either, since must be tackled the computational problem of identifying homology between sequences that can be built of gain, loss and rearrangements of domains. This is why this strategy combines architectures of canonical domains of IIS genes for an efficient detection of ultraconserved domains between the study species. The strategy is designed to identify tracts of domains in tunicated species that lack of transcriptomics or proteomics data. Therefore we propose a model to identify possible putative genomic regions associated with coding for IIS domains in the draft genome of a species without annotation and expression data such as the tunicate *Didemnum vexillum*.

Finally, an evolutionary model of gain and loss of ultraconserved domains of putative ISS genes is implemented. This pipeline was applied to all the genomes of five species of tunicates and of an external group consisting of other chordates.

The characteristics of the genomes evaluated during this thesis, especially in tunicates, represented three types of important computational challenges. First genomes with peculiar evolutionary histories, second for some of these species the assembled genomes are highly fragmented and since they are non-model organisms they do not have extensive experimental information that allows us to train and use gene annotation programs widely used in chordates as the Ensembl pipeline and third there is complexity in the gene architecture of the genes of the IS since they present duplications of domains, rearrangements and losses.

These problems were solved in Chapter 1 through an analysis focused in the wide repertoire of the existing gene architecture in two main databases `InnateDB` and `Insect Innate Immunity Database` (IIID) that let us to define a system of "Gold Standard"domains which

was mapped into species articulated in the Ensembl using BioMart, which included the species `Ciona intestinalis, Ciona savignyi, Petromyzon marinus, Latimeria chalumnae and Danio rerio` to identify a set of canonical architectures of the IIS of lower chordates. Later on, were used protein sequences annotated by other systems as targets for the identification of canonical domains associated with the "Gold Standard"previously established for the species of tunicates `Oikopleura dioica and Botryllus schlosseri` and the `protocordado Branchiostoma floridaeo`. In chapter 2 we present the strategy used to identify domains in species that lack of experimental evidence of expression and annotation such as the tunicate *D. vexillum*. This restriction in the number of evaluated domains allowed quickly, accurately and efficiently to establish sets of putative domains of protein architectures similar to those reported in the literature. This approach could be used as the starting point for the search of homology relationships, mainly of orthology and paralogy and as a model of gains and losses of ultraconserved ISS domains as is described in the Chapter 3.

Key words: Domains, Inmune System, Innate Inmune System, Tunicates, Gene annotation, Gain and losses.

Contenido

Lista de Tablas

Lista de Figuras

Objetivos

Objetivo general

Construir procesos computacionales automatizados tipo *pipelines* que permitan identificar la organización genómica de los dominios ancestrales asociados al sistema inmune en cordados basales, para elaborar un modelo básico que explique la pérdida, ganancia y evolución de dichos elementos.

Objetivos específicos

- Diseño e implementación de una *pipeline* que automatice el flujo de procesos asociados a la identificación de exones, e identificaciones de posibles dominios asociados al sistema inmune basados en una aproximación de organización de exones en clusters y su posterior validez usando modelos ocultos de markov.
- Construcción de una *pipeline* que integre los procesos para la detección de ortología, perdida y ganancia de dominios con el fin proponer un modelo de evolución de dominios del sistema inmune en las especies de interés.
- Generar una estrategia computacional de anotación de algunos genes asociados al sistema inmune y de sus regiones adyacentes, para presentar un primer borrador del inmunogenoma de *Didenmum vexillum* .

Introducción

Generalidades

Posterior al secuenciamiento y ensamble de un genoma, los investigadores se enfrentan con dos retos computacionales adicionales que consisten en la detección de las subregiones genómicas que codifican en una primera instancia y en la otra en la asignación de su posible funcionalidad. Estos dos pasos principales conocidos como la predicción y anotación de genes, términos que han sido usualmente usados por algunos autores como sinónimos, son procesos diferentes, es decir, los predictores de genes se encaminan a la identificación de la secuencia codificante (CDS) del inglés coding sequence (CDS) mas probable que excluye en primer lugar las regiones no traducidas del inglés (UTRs) y en segundo lugar las variantes en las secuencias que definen los limites de los exones o la heterogeneidad de los mismos que pueden detectarse por la evidencia de transcritos alternativos. Esta amplia complejidad del funcionamiento del gen eucariote, que depende altamente de los motivos de subsecuencias en los genomas, son incorporados principalmente en el proceso de anotación de genes. Hoy en día la anotación se ha convertido en una tarea mucho más compleja que la predicción de genes [121]. Aunque en sus primeros inicios la anotación se basó en la búsquedas de señales propias de los genes eucariotas, como búsquedas de señales de cajas TATA, de definiciones exo/intrónicas y señales de poliadenilación, en la era post-genómica, la anotación de genes se ha abordado de dos formas: procesos manuales mediados por curación manual como el proyecto VEGA-HAVANA del Wellcome Trust Sanger Institute o automatizados como las pipelines de Ensembl. La diferencia entre ellos es la rapidez y eficiencia en las cuales ocurren los dos procesos y la calidad de las predicciones de ambos modelos, siendo más eficiente el proceso automatizado pero más preciso el proceso manual. Entre algunas de las bases de datos que integran pipeline automatizadas se encuentran adicional al Ensembl, FlyBase y otros como Biopipe anclado en Open Bioinformatics Foundation (OBF; http://www.open-bio.org), Genescript y ASAP [92].

El reto de poder caracterizar de forma rápida y confiable las secuencias provenientes del secuenciamiento de nuevos genomas de forma masiva, toma mayor importancia debido a la creciente necesidad de poder extraer la máxima información y poderla interpretar de acuerdo a las relaciones de homología, funciones celulares y aproximaciones bioquímicas, siendo el fin último de la anotación, otorgarle una función biológica a una secuencia genómica que pueda ser usada en estudios funcionales o evolutivos [1] [13] [6] [109] [110].

Si es claro que la fisiología celular es compleja y que la regulación de expresión de genes

a nivel de la transcripción y del procesamiento de mRNAs genera formas alternativas entonces por ende si la definición de lo que una UTR o un Exón son es compleja, entonces se hace mucho mas difícil la definición de lo que una región genómica asignada a un gen pueda ser. Sin embargo, actualmente existen pipelines que pueden manejar la heterogeneidad de los datos de expresión como EST o RNA-seq u homologias de proteinas integradas con modelos coherentes de genes conocidos para caracterizar todos los elementos funcionales que componen los genes, finalizando con la asignación de una función biológica a una secuencia genómica. [1][13]. Para el caso de los vertebrados y especies cercanas, como algunos de los genomas utilizados en este trabajo, éstos han sido sometidos a un riguroso proceso de anotación ampliamente aceptado en la genómica comparativa de vertebrados como es la pipeline de anotación del laboratorio de biología molecular europeo (EMBL) que integra la información experimental de 70 especies de vertebrados y de otras especies de metazoarios bajo la plataforma Ensembl [1].

Dentro del grupo de especies trabajadas nos concentraremos en algunas especies del subphylum Tunicata, este grupo fue de nuestro interés por tres razones: 1) Al ser el grupo hermano de los vertebrados los provee de una posición filogenia privilegiada 2) esta extraordinaria ubicación a su vez los convierte en un grupo clave en el estudio de la evolución del sistema inmune, ya que se encuentran justo antes del fenómeno biológico que dio origen a la complejidad atribuida al SIA, el BigBang inmunologico [46] 3) Este grupo se caracteriza por la gran diversidad de formas de vida, nichos ecológicos y ambientes marinos [21, 69].

Estos tres atributos postulan a los tunicados como un grupo modelo que puede llegar a explicar las estrategias de en respuesta inmune innata que les han permitido sobrevivir en los diferentes hábitats según sus diversas formas de vida (solitarios, sociales y coloniales). En conjunto, se cree que pueden tener relaciones complejas entre el ambiente, el genoma y el fenotipo y es allí en donde se espera diversidad en la composición de sus sistemas inmunes.[28, 10].

Sin embargo pese a la importancia de estas especies en la integración de disciplinas de la biología y la biología evolutiva del desarrollo los estudios genómicos y comparativos son escasos. Hasta el momento se han anotado los genomas de cuatro ascidias, tres solitarias: *C. savignyi*, *C. intestinalis* [33, 105] y *O. dioica* [34, 104] y una colonial:*B. schlosseri* [116]. Estos genomas han permitido estudios de genómica comparativa entre algunos tunicados con genomas de algunos cordados que han perimido identificar en estos últimos fenómenos de expansión de familias genicas por eventos de duplicaciones locales, pero también perdida masiva de genes.

Estos genomas muestran una complejidad en la organización genómica con altas tasas de evolución que ha permitido inferir la existencia de un procesos de re-estructuración genómica que han conducido a una organización genómica particular algunos tunicados, pero aun no esta claro si este proceso es exclusivo de algunas especies o es común a todos los tunicados [94, 10].

Estos eventos de re-estructuración también se observa en sistemas complejos que mantienen

la integridad de los organimos como el sistema inmune, un sistema que como demostraremos más adelante ha sido fundamental en la evolución de la vida multicelular y debido a su complejidad y diversidad en su evolución ha sido objeto de esta tesis.
La importancia del estudio del sistema inmune radica en que este es mucho más que un sistema de reconocimiento y eliminación de patógenos, es un sistema organizado por módulos los cuales al interactuar entre ellos permiten mantener la integridad genética, fisiológica y el equilibrio homeostático en un organismo. Dichas interacciones modulares y los módulos en sí mismos son moldeados por los planes corporales, historia y esperanza de vida de cada organismo.

Algunos estudios rastrean el origen del sistema inmune más allá de 500 millones de años, es por ello que no pasa desapercibido la presencia de estrategias de reparación de tejidos así como de reconocimiento, neutralización, eliminación de patógenos, rastreada en ramas comunes de la evolución de los grupos animales, como es el ejemplo del mantenimiento de la integridad individual del organismo por medio del reconocimiento de lo propio de lo extraño o aloreconocimiento; se ha propuesto que la selección de este mecanismo permitió la existencia de organismos multicelulares que fue derivando en la capacidad de evitar la formación de quimeras. [22] [84] [5].
De igual forma, el sistema inmune también ha derivado en mecanismos que son innovaciones evolutivas de ciertas ramas en metazoarios. Como es el caso de del sistema inmune específico, adquirido, anticipatorio o adaptativo (SIA) el cual aparece como una innovación en los vertebrados mandibulados y cuyo mecanismo se basa en la expansión de células clonales especializadas (linfocitos). Los linfocitos poseen receptores (Anticuerpos y receptores de células (TCR)) con una variabilidad casi ilimitada generada por recombinación somática mediada por enzimas tipo RAG (dicha recombinación le permite a los vertebrados mandibulados el reconocimiento de antígenos en constante cambio). [7] [56]. Por el contrario, el sistema inmune no anticipatorio o innato (SII), es un claro ejemplo de estrategias conservadas a lo largo de la evolución de los animales, esto se debe a que el SII se basa en el reconocimiento de patrones moleculares, que pueden ser de tipo microbiano y que incluyen polisacáridos complejos, glicolípidos, lipoproteínas y que comparten la característica de estar presente en el parásito más no en el hospedero conocidas como los Patrónes de Moléculas Asociadas al Patógeno (PMAP).
Existen muchas preguntas abiertas sobre la evolución del SI por ejemplo: ¿Cómo son las dinámicas evolutivas que permiten el surgimiento de innovaciones especie específicas como el SIA? o por el contrario ¿Qué permite la fijación de estrategias conservadas como el SII? todas ellas siendo hoy en día un tema que necesita una investigación a fondo Algunos estudios han propuesto que en términos de ancestralidad, la inmunidad se puede rastrear desde las plantas, protozoos hasta bacterias, y se han propuesto 5 hitos fundamentales de la evolución de este sistema: 1) la aparición de componentes del sistema complemento, moléculas Anti-trematodos encontradas en *Lophotrochozoa* hace aproximadamente 520- 530 millones de años, péptidos Antimicrobianos en nudibranquios (*Mytilus galloprovincialis*, *Biomphalaria*

glabrata), 2) la Óxido nítrico sintasa (NOS) encontrada en *B. glabrata* y en *Schistosoma mansoni* , que sirve de defensa, ya que el óxido nítrico y el producto de su reacción con superóxido (peroxinitrito) puede causar daño a las estructuras proteicas, 3) el estallido respiratorio detectado en moluscos como en (*Lymnaea stagnalis* , *S. mansoni* , *Crassostrea gigas* y *M. galloprovincialis*), que es el proceso mediante el cual se genera una cascada inicializada por el complejo enzimático NADPH oxidasa, provocando la producción de anión superóxido y radicales los cuales son perjudiciales para microorganismos, 4) los patrones de reconocimiento PRR encontrados en *Helix pomatia*, *B. glabrata*, que son receptores encargados del reconocimiento de patrones PMAP. Los patrones PMAP aunque amplios son limitados, son compartidos por diferentes parásitos y son conservadas entre organismos y fundamentales para la supervivencia del parásito. [7] [56] *Helix pomatia*, *B. glabrata*[9]; por ultimo 5) El surgimiento del SIA[22].

Este último hito, es conocido como el Big-Bang inmunológico, cuyo origen fue el resultado de dos rondas de duplicación génica, de las cuales no se tiene claro si fueron locales o a lo largo de todo el genoma, pero que dieron lugar a expansión de genes basados en dominios tipo inmunoglobulina. Dichas duplicaciones tuvieron lugar después de la separación del linaje que conduce a los Cefalocordados [58]. A partir de este Big-Bang se generaron dos eventos que tuvieron un carácter importante en la aparición de sistema inmune adaptativo, 1) la adquisición de los genes RAG y 2) la recombinación VDJ en los linfocitos. Estos dos eventos sólo han podido ser rastreados hasta los peces mandibulados, siendo totalmente ausente en el grupo hermano de los vertebrados, los urocordados [45]. Aunque no es rastreable el sistema inmune adaptativo en los tunicados, hay evidencias que sugieren que siendo los tunicados el grupo ubicado en tiempo evolutivo inmediatamente anterior a la aparición de la inmunidad adaptativa, puedan existir evidencias de un fenomeno evolutivo que implica el uso de nuevas funciones de células ancestrales involucradas en el alorreconocimiento que cumplen el mismo papel que las células NK, con arquitecturas similares a los receptores de estas células en los vertebrados mandibulados, dominios de lectina tipo C; moléculas con un ITIM y un inmuno-receptor basado en tirosina; y la activación del motivo (ITAM) [75] [35] . Otro punto importante que hay que tocar con respecto al sistema inmune en vertebrados no mandibulados y conservado en los mandibulados, es el sistema del complemento, el cual cumple tres funciones importantes: opsonización de las partículas extrañas, inducción de reacciones inflamatorias y citólisis. De las moléculas conservadas del sistema del complemento las más representativas son el C3 y el BF (factor B) [83].

Es por eso que el estudio de los tunicados, un grupo hermano de los vertebrados mandibulados que al encontrarse en la frontera de la emergencia de la inmunidad adaptativa, se presenta como un grupo clave para entender el surgimiento de la complejidad y de la diversidad del sistema inmune adaptativo. Es por ello que entender las dinámicas de este grupo posibilita preguntas referentes a entender la transición de un sistema con la presencia exclusiva de la inmunidad innata a sistema que cuentan con ambas[71]. El ejemplo más claro de dicha transición, es la presencia de una región proto-MHC en los cefalocordados *Brachiostoma*

floridae y *Brachiostoma belcheri* los cuales poseen pocos homólogos con los genomas de los tunicados disponibles[56]. La importancia de esta proto-región está dada a que el MHC es un complejo exclusivo de la inmunidad adaptativa y es clave para el proceso de reconocimiento de lo propio y de lo extraño en los vertebrados mandibulados y está completamente ausente a nivel funcioinal en los invertebrados y tunicados. Dicha región precursora del MHC muestra una alta tasa media de reordenamiento local y se cree que es uno de los puntos de origen del big bang, en donde muchos nuevos dominios y combinaciones de dominios surgieron en esta región, y contribuyeron al origen de la inmunidad adaptativa [7] [58]. La hipótesis del reordenamiento de dominios y exones pueden haber sido una fuerza impulsora de la evolución del sistema inmune, no solo aplica al surgimiento del SIA. Paalsson, et al. 2007 demostró a través de un estudio de homologías estructurales, que la totalidad del sistema inmune ha estado basado fundamentalmente en nueve dominios ancestrales los cuales son conservados a lo largo de la evolución de múltiples organismos, sugiriendo que el sistema inmune proviene de unas pocas proteínas ancestrales, seleccionadas y fijadas a lo largo de la evolución [88]. Siguiendo esta idea asi como ideas similares discutidas por otros autores fundamentamos la manera como se abordó esta tesis, es decir un trabajo basado en el estudio del dominio proteico como el módulo evolutivo más elemental del sistema inmune y por ende de manera preliminar presentamos dinámicas de ganancia y perdida de dominios así como combinaciones que puedan ser consideradas como un resultados del efecto de una fuerza impulsora evolutiva que conllevó a la complejidad del sistema inmune. Teniendo entonces en cuenta que la variabilidad de dominios implica a escala exónica organizaciones nuevas (sin tener en cuenta que existe también el reordenamiento de exones), entonces, es sumamente interesante describir si la variabilidad de dominios hubiese favorecido la creación de nuevas arquitecturas proteicas que pudieron ser seleccionadas positivamente [88].

Modulos del Sistema Inmune

Moléculas de reconocimiento

Entre las proteínas asociada al sistema inmune de invertebrados se encuentran los receptores encargados del reconocimiento de moléculas constitutivas de microorganismos como el peptidoglicano, la flagelina o el lipopolisacárido (LPS) exclusivo de bacterias Gram-negativas y los ácidos teicoicos y lipo-teicoicos (LTA) exclusivos de bacterias Gram-positivas , ARN de doble cadena de virus y beta-glucanos de hongos, entre otras moléculas consideradas patrones moleculares asociados a microbios (MAMPs, del inglés Microbe Associated Molecular Patterns), o muestran daños internos por medio de las proteínas alarminas [115] [98] [101] [12]. Estos receptores se caracterizan por tener dominios LRR (Leucin Rich Repeat), ig o dominios de lectina [103]. Entre ellos el más destacado es el TLR o (Toll-Like Receptors), cuya arquitectura canónica se caracteriza por tener dominios LRR + TIR (receptor de interleukina-1) [4] [76]. A lo largo de la evolución se han encontrado formas alternativas

de TLR como en Hydra en donde los dominios TIR y LRR están separados en dos moléculas HyLRR-2 y HyTRR-1. Estos dominios LRR son también componentes de los receptores NLR (NOD-like receptors), los cuales tienen tanto un dominio de unión a nucleótido como los dominios (NACHT) y un dominio encargado de la señalización (CARD)[12] [67]. Por otro lado también se han encontrado receptores de interleucinas que sus componentes son TIR + dominios de inmunoglobulina extracelulares [79] . Por otro lado, los dominios de inmunoglobulina propician la interacción proteína-ligando dando lugar al reconocimiento específico o genérico de diversas moléculas [30]. Otro tipo de dominio constante en la evolución del sistema inmune son los dominios tipo lectina, los cuales generalmente son dominios de unión a carbohidratos, presentes en los patrones moleculares como en las envolturas celulares de bacterias y hongos [60], las más representativas de este grupo son las galectinas, las lectinas tipo C, las pentraxinas, las ficolinas y las taquilectinas. [60] [32] [73]. Las lectinas tipo C son dependientes de calcio y normalmente están ancladas a la membrana exceptuando las colectinas [26]. Otro grupo son las pentraxinas que tienen una estructura cíclica multimérica involucradas en respuesta al daño tisular, en el proceso inflamatorio y en la infección.
Por otro lado se encuentran las proteínas de unión a lipopolisacárido de bacterias Gram-positivas y Gram-negativas que poseen dominios BPI/LBP/PLUNC, y en mamíferos está relacionado con el procesamiento de lípidos para ser reconocidos por los TLR[102] , otro grupo de receptores son los scavenger de naturaleza transmembrana y se unen a lipoproteínas de baja densidad modificadas en bacterias [4] [80]. Otro tipo de moleculas de interes para comprender el funcionamiento del sistema inmune son las moléculas de adhesión como las caderinas y las selectinas. Las cadherinas están presente en uniones adherentes entre células, son de origen transmembranal y poseen un número variable de dominios de cadherina extracelular a nivel citoplasmático está conectada con el citoesqueleto de actina [85]. Las selectinas son lectinas tipo C transmembranales especializadas en el reconocimiento de carbohidratos propios y en mamíferos están asociadas a un papel importante en la vigilancia de células inmunes [123] [60] [26]. En mamíferos, las selectinas son receptores endoteliales que juegan un papel importante en la vigilancia de células inmunes, no detecta patógenos pero mantiene la integridad de los tejidos y posee la capacidad de interacción con otras proteínas endógenas y exógenas [123]. En mamíferos, las moléculas con repeticiones de trombospondina tipo 1 (TSP1) [103] mediante la fijación celular, se unen a glicosaminoglicano, inhiben la angiogénesis, activan la vía de señalización del factor de Necrosis Tumoral Beta (TGFß) e inhiben metaloproteinasas de la matriz extracelular[108].

Modulos de señalización

Después de la detección del patógeno son activadas rutas de señalización citoplasmáticas que activan la expresión de genes que pueden terminar favoreciendo sus productos procesos inflamatorio mediado por moléculas efectoras o en la muerte celular con el proceso de apoptosis [79]. En el caso de los TLR el adaptador que desencadena la respuesta es el MyD88

en conjunto con proteínas reclutadas como RAK, TRAF y TAK1, propiciando la expresión de genes efectores y pro-inflamatorios mediados por factores de transcripción NF-k beta [47] [86]; estos adaptadores de vías de señalización citoplasmática contienen dominios DEATH, MyD88 y el factor de transcripción NF-k beta [79]. Otras moléculas involucradas en el proceso apoptótico son MAPK a través de la señalización por FGF [97] , Wnt [81] , Notch, Hedgehog y PI3K.

Sinergimos entre dominios asociados al Sistema Inmune

El sistema inmune en los cordados basales como en los tunicados tienen al igual que en los vertebrados superiores elementos estructurales ancestrales que están presentes en los protocordados; dichos elementos estructurales son los dominios de Ig, dominios de lectina y los dominios ricos en repeticiones leucina (LRR). Partiendo de los mismos componentes que en vertebrados superiores, se ha planteado la posibilidad que los vertebrados sin mandíbula puedan tener un sistema inmune adaptativo de tipo .ªnticipatorio", ya que poseen una alta diversidad de receptores carentes de recombinación mediada por las RAG [55] . Dicha recombinación es compensada por una reorganización de genes codificantes para dominios ricos en LRR, que además es enriquecida por mecanismos como la duplicación de genes; deleciones y recombinaciones; familias multigénicas; variación en el número de copias génicas; diversificación alélica; splicing y edición de RNA. Esta diversificación es evidente en genes como los VLR [72] [55] [16] [17], y en receptores VCBP de anfioxo que comparten diferente origen, y sin embargo en la misma proteína se encuentran genes de Ig y lectina (VCBP). Otro tipo de receptor típico de los vertebrados inferiores es el receptor tipo peaje (TLR) que ayuda a la defensa del huésped con la expresión de genes que no pertenecen al MHC-I conocidos como genes de armazón [78] [117] . Estos TLR son proteínas transmembranales que poseen dominios TIR en el interior de la célula; dominios ricos en motivos de LRR al exterior y su principal característica es el reconocimiento de PRR y son de vital importancia en el asentamiento de las colonias en su respectivo sustrato ya que se reconoce la flora bacteriana favorable a dicho asentamiento [83]. En algunos tunicados, ya sean solitarios se conocen que existen células ancestrales involucradas en el alorreconocimiento que cumplen el mismo papel que las células NK, con arquitecturas similares a los receptores de estas células de vertebrados superiores, dominios de lectina tipo C; moléculas con un ITIM y un inmuno-receptor basado en tirosina; y la activación del motivo (ITAM) [75] [35].

Desde el punto de vista evolutivo, en especies como el erizo de mar y otras especies de invertebrados, se puede observar que a lo largo de la evolución del sistema inmune hay mecanismos de reconocimiento conservados que son moldeados por los planes corporales, historia y esperanza de vida de los organismos y que estos fenomenos generan variabilidad y plasticidad en receptores del sistema inmune que en algunos casos puede ser rastreada en ramas comunes de la evolución o por el contrario terminar en una rama única para cada especie. Ahora, si el sistema inmune es mucho más que reconocimiento y eliminación de patógenos y es el

encargado de mantener la integridad genética, fisiológica y el equilibrio homeostático en un organismo [22] entonces hace apasionante la idea de construir modelos y flujos automatizados para estudiar el sistema inmune ya que en si mismo, por su historia evolutiva y complejidad, plantea un reto computacional para estudiar la variedad de estructuras y riqueza de sus dominios, lo cual desafia no solo los principios basicos de la biologia sino también al campo de la anotación genómica en bioinformática ya que búsquedas convencionales de homología no pueden ser aplicadas directamente en este tipo de especies debido a que estudios de seleccion positiva indican que al contrario del resto de proteinas, en el sistema inmune la seleccion natural no favorece la baja variablilidad en las estructuras sino que al contrario permite y favorece la plasticidad del sistema inmune puesto que de esa plasticidad depende el éxito del mismo planteado claramente por Buckley y Rast (2015) en su discusión sobre la diversidad de receptores del sistema inmune en animales y los orígenes de la complejidad en los deutorostomados.

Ahora el hecho de ser el sistema inmune de los tunicados el mas inmediato antes de la expansión de las inmunoglobulinas presentes en cordados, es de gran interés caracterizar este sistema en este grupo por estar en la frontera de la emergencia de la inmunidad adaptativa, pero dada la complejidad de organización de sus genomas, heterogeneidad en los rearreglos genómicos y altas tasas de cambio, el proceso computacional requerido para anotar estos genes en estos genomas no es sencillo y depende altamente de la complejidad y arquitectura de los genes asociados al sistema inmune como se aborda en esta tesis y del estado de los ensambles disponibles. Es bien conocido que en estos genes la diversidad de las arquitecturas implica combinación, rearreglos de dominios, copias y perdidas de dominios generando un reto en el computo. Por otro lado, el grupo de estudio escogido en esta tesis es grupo hermano de los vertebrados y un grupo modelo para estudiar la evolución del sistema inmune ya que se encuentran previos a lo conocido como el bin bag inmunológico. Parte de los datos utilizados en está tesis están basados en los dominios de los receptores asociados al sistema inmune innato de arquitecturas canónicas conocidas. Esta estrategia alternativa se propuso, ya que la búsqueda por medio de métodos computacionales tradicionales como BLAST de secuencias primarias altamente divergentes pueden quedar afuera del análisis, mientras que el estudio por medio de arquitecturas estereotipadas de dominios o arquitecturas canónicas son un método bastante sensible para identificar homólogos muy distantes puesto que se basa en la función que la arquitectura generada desde receptor se produce y no es basado directamente en la homología única de la secuencia nucleotídica como lo propone Buckley y Rast (2015) [22].

Contexto Computacional

Anotación génica

Uno de los retos computacionales de la última década en genómica ha sido poder resolver la anotación automatizada a gran escala de los genomas logrando que se que cumpla con estándares de velocidad, eficacia y consistencia en las anotaciones para asi enlazar una secuencia genómica con una función biológica [1], [13] Dos métodos principales se han utilizado para lograr anotar los genes: los primeros llamados métodos *ab initio* y los métodos basados en datos de expresión o datos funcionales ya sean EST, cDNAs, datos de RNAseq o proteómicos.

Métodos ab initio

Los métodos ab initio fundamentan sus aproximaciones en algoritmos que capturan información dependiente de señales típicas que definen los genes procariotes y eucariotes. En éstas se encuentran la posibilidad del método en predecir ORF (open reading frames) para los cuales codones de inicio y de parada deben ser detectados, señales de los intrones para los eucariotes y definiciones de UTRs y regiones reguladoras de la transcripción como cajas TATA y señales de poliadenilación en los extremos 3' de los genes. Aunque desde los años 90s existían muchos predictores basados en búsquedas de homologías o en campos azaroso condicionales de modelos de genes bacterianos como las usados en GLIMMER `http://www.cs.jhu.edu/~genomics/Glimmer/` para una mejor comprensión del estado del arte para esa época se puede revisar en la publicación de Claverie en 1997 [29], se resalta principalmente en este trabajo los predictores de genes basados en Módelos de Markov. Caveliere resalta que "en estos modelos las secuencias biológicas son modeladas como salida de un proceso estocástico en el cual la probabilidad para que un nucleótido dado ocurra en una posición p depende de la posición ocupada en k previas posiciones. Tal representación es llamada un modelo de Markov de orden k, entonces diferentes señales en la secuencia de los genes pueden exhibir diferente periodicidad y por tanto corresponden a diferentes modelos de Markov" Algunos modelos como GenMark `http://exon.gatech.edu/GeneMark/` se basan en determinar si para una región de DNA dada es mas probable encontrar un modelo de Markov codificante en comparación con un Módelos de Markov no codificante que es construido por entrenamiento previo los cuales hoy en día fueron extendidos de su aplicación sobre genomas bacterianos a modelos en eucariotes con extensiones utilizando heurísticas, métodos de entrenamiento no supervisado y métodos de entrenamiento semi-supervisados usando datos de RNAseq como GeneMart-ET. Extensiones a las primeras aplicaciones de los Modelos de Markov fueron los modelos ocultos de Markov (HMMs) en los que las secuencias se representan según Caveliere como "una salida de un proceso abstracto que avanza a través de una serie discreta de estados algunos de los cuales son ocultos al observador". Estos modelos fueron ampliados a un nivel más alto de abstracción en los modelos gene-

ralizados ocultos de Markov en donde los estados son sub-modelos (redes neuronales, matrices de posiciones) que dan como salidas estados como los implementados en GENSCAN `http://genes.mit.edu/GENSCAN.html`. En esta herramienta se introdujo originalmente un modelo probabilístico para la predicción de estructura del genes de humano, como las características de distribuciones de composiciones de exones, intrones y regiones intergenicas que les permiten una buena predicción de genes [23]. Uno de los métodos establecidos para la predicción de genes ab initio en el 2005 generó un gran avance en los procesos de anotación como la herramienta AUGUSTUS [106] que como método ab initio permite con bastante exactitud la predicción de genes y hoy por hoy existen versiones que pueden ser entrenadas integrando datos de RNAseq, cDNAs cortos, lecturas cortas sencillas o pareadas. AUGUSTUS corresponde a la serie de modelos Generalizados ocultos de Markov que usa las señales contenidas en las secuencias genómicas y que puede también usar información extrinseca en algunas de sus extensiones. Otros métodos utilizados en la predicción de genes basan sus aproximaciones en sistemas basado en reglas como lo hace GeneID `http://genome.crg.es/geneid.html` que usa un método heurístico basado en reglas que permiten ensamblar las señales típicas de los genes en un producto o modelo mas probable y que puede ser el asignado como gen, otros métodos son basados en linguística como GenLang, o en LDA (linear discriminant analysis) como FGENEH o en árboles de decisión como en la aplicación MORGAN o herramientas que utilizan los principios de programación dinámica como GENVIEW[29] .

Métodos híbridos basados en expresión

Los problemas de la anotación genómica basada en datos de expresión se pueden dividir en dos perspectivas: primero la centralización de la información derivada de experimentos en animales modelo como los vertebrados y segundo la manera como se afronta la anotación desde un esquema manual o automatizado.

Tanto la centralización de la información como el tipo de aproximación afrontan los retos propios de trabajar con los datos masivos derivados de las nuevas tecnologías de secuenciamiento como la complejidad de la asignación funcional y codificante, asi como afrontar el incremento de la información a analizar dado el incrementos de secuenciación de genomas complejos que no son fáciles de ensamblar.

El primer problema al que se enfrenta el investigador hoy en día en la anotación, es la alta dependencia de pipelines basadas en datos de expresión de experimentos en humanos o de animales modelos como el ratón o zebrafish sesgando sus resultados y generando como consecuencia que un problema en la anotación de los genomas de especies no modelos.

Uno de los problemas con las especies no modelo es que la información disponible no cuenta con una relación evolutiva con la especie de estudio, es el caso de nuestro grupo de estudio los tunicados donde las características evolutivas difieren enormemente de las especies modelo, sumado a que han sido poco estudiado y el acceso a datos experimentales es casi nulo, existe una excepción en Ciona intestinalis, pero al ser altamente divergente de las especies

coloniales presenta los mismos retos computacionales que el resto de especies modelo. [13].

El segundo problema en cuanto a cómo se ha hecho la anotación en la última década, se destaca la anotación manual, la cual tiene como gran ventaja producir datos de más alta calidad y estructuras de genes más precisas ya que normalmente es dada por un equipo de anotadores altamente entrenados como ocurre con el proyecto VEGA-HAVANA del Wellcome Trust Sanger Institute `http://www.sanger.ac.uk/science/tools/vega-genome-browser`. Por otro lado, la predicción totalmente automatizada de las estructuras de genes tiene la ventaja de ser rápida, no requiere un equipo humano de anotadores manuales entrenados, y procesa el análisis en bruto de forma consistente como lo hace por ejemplo MAKER [27]. Aunque estos métodos pueden considerarse menos sesgados por la influencia del criterio humano, sus pipelines de anotación generan falsos negativos que conllevaron en el pasado a una baja especificidad en sus aproximaciones [13] pero que pueden ser posteriormente corregidas por un proceso de postprocesamiento manual.
Sin embargo, con el aumentado el número de genomas secuenciados, pipelines clásicas como las utilizadas por el Ensembl desde el año 2016 `http://www.ensembl.org/info/genome/genebuild/genome_annotation.html` están adaptando sus aproximaciones a las nuevas exigencias de la secuenciación másiva de genomas, ya que están diseñadas para integrar la anotación de forma independiente, es decir, de los genomas ensamblados borrador altamente fragmentados y de la información limitada de secuencias proteicas, RNA-seq o de cDNA de secuencia completa, [1]. Sin embargo, pipelines como AGOUTI son disponibles hoy en dia para resolver de forma eficiente el problema de anotar sobre genomas altamente fragmentados puesto que usa datos del transcriptoma y RNA-seq, pero a diferencia de Ensembl, esta pipeline ayuda a reorganizar el ensamblaje de contigs a scafolds como una estrategia para evitar la redundancia de información que por supuesto genera el alto número de fragmentos genómicos [124]. A pesar de haber superado algunos límites clásicos de las pipelines de anotación, AGOUTI permite identificar un conjunto de señales de motivos propios de un gen aun no es de todo eficiente para predecir genes sobre genomas ultrafragmentados[1].
Por último el secuenciamiento de nueva generación basado en lecturas cortas, ha dado la posibilidad de secuenciar nuevos genomas a bajo costo y por ende incrementa la mayor cantidad de genomas disponibles para analizar, pero presenta el inconveniente de que estas estrategias de secuenciamiento limitan el proceso del ensamblaje de los genomas y por ende quedando muchos de ellos en genomas muy fragmentados, lo cual trae como consecuencia dos cosas, que se pierda la vecindad y colinearidad de los genes y que un gen sea ambiguamente ubicado en más de un fragmento genómico, imposibilitando la identificación correcta del gen y generando un sesgo en el número e identidad de los genes hallados en una especie. Pese a los retos que conlleva la anotación como lo hemos mencionado anteriormente, es novedosa la pregunta bioinformática que intenta responder esta tesis, ya que se plantea una estrategia de computo que se basa primero en una buena definición de arquitecturas conocidas del sistema inmune candidato y en una refinada definición de un conjunto Gold Standard de dominios

específicos de tunicados y cordados basales que han sido curados y preliminarmente anotados en el Ensembl.

Lista de símbolos

Hola

Símbolos con letras latinas

Símbolo	Término	Unidad SI	Definición
A	Área	m^2	$\int \int dxdy$

Símbolos con letras griegas

Símbolo	Término	Unidad SI	Definición
α_{BET}	Factor de superficie	$\frac{\mathrm{m}^2}{\mathrm{g}}$	$(w_{\mathrm{F,waf}})(A_{\mathrm{BET}})$

Subíndices

Subíndice	Término
0	Estado de referencia

Superíndices

Superíndice	Término
n	Coeficiente x

Abreviaturas

Abreviatura	**Término**
SI	Sistema Inmune
Ciin	*Ciona intestinalis*
Cisa	*Ciona savigny*
Dive	*Didendum vexillum*
GO	Gene Ontology

1 Capítulo 1: Construcción de Pipeline automatizada para la identificación de dominios asociados a proteínas del sistema inmune de Tunicados

La anotación *de novo* e identificación de genes es una necesidad bioinformática actual debido al auge del secuenciamiento de nuevos genomas, que durante las últimas décadas ha generado información genómica de especies modelo y de especies no modelo como algunas de las especies trabajadas en esta tesis y que son de gran importancia para comprender la evolución de los vertebrados o especies clave para entender la evolución de sistemas complejos como el sistema inmune. Es claro que la asignación de posibles funciones biológicas a los genes identificados es un reto computacional pero es uno de los pasos a seguir mejorando en métodos bioinformáticos como experimentales en la era post genómica. Este desarrollo implica la creación de flujos de trabajo que integren herramientas especializadas con el fin de lograr un análisis automatizado que sobre dichos genomas de forma más eficiente permitan una anotación génica adecuada [92, 1]. Aunque existen pipelines ampliamente usadas, muchas de ellas son dependientes en gran medida de la información experimental disponible ya sea de datos de RNAseq, cDNA y secuencias proteicas. Estos datos facilitan la anotación de genes en organismos modelos cercanos evolutivamente pero en ciertos grupos de organismos que no son modelo dicha información no crece en las misma proporción o es ausente en comparación con la cantidad de genomas disponibles. Este auge en el secuenciamiento de genomas se suma a la baja resolución y fragmentación del ensamblaje de muchos de estos genomas lo que hace aun mas difícil el uso de pipelines convencionales.

Entonces se hace necesario desarrollar nuevas opciones de anotación y asignación de funciones para los organismos no modelo que no dependan altamente de información experimental y que sean flexibles para ser usadas sobre genomas no completamente ensamblados a nivel de cromosomas. Entre otras opciones, una aproximación ampliamente usada ha sido la aplicación de los Modelos ocultos de Markov (HMM, por sus siglas en inglés Hidden Markov Model). Los HMM han sido ampliamente usados en biología para la identificación de las subunidades de las proteínas conocidas como dominios proteicos y que son a la vez la unidad fundamental de la arquitectura funcional de una proteína. Por tanto su identificación permite aproximarse a la arquitectura de la proteína y por ende proponer una función biológica. Los

HMM al estar basados en un modelo probabilístico enriquecido con una gran cantidad de alineamientos de secuencias de muchas fuentes de diversas proteínas de múltiples especies, son modelos de referencia bastante idóneos para ser aplicados en la identificación de regiones genómicas candidatas a codificar para proteínas no solo para especies con poca información de referencia como ocurre en los tunicados sino para grupos de proteínas pertenecientes a sistemas complejos como lo es el sistema inmune innato o el adaptativo (SII y SIA).

Debido a que los tunicados se encuentran en la frontera evolutiva de especies que carecen de un SIA pero preceden al grupo de especies en donde aparece y sumado a evidencias que denotan puntos donde pudo emerger el SIA como ocurre en el cefalocordado B. floridae y su región no funcional conocida como Región proto-MHC, se han convertido los tunicados en un grupo clave para entender la complejidad del Sistema inmune. . Este grupo está ubicado previa a las dos rondas de duplicación del genoma reportadas posterior a la aparición de los urocordados y previa a la radiación de los vertebrados mandibulados [74].. Por tanto es de esperarse que existan huellas en sus genomas de posible nacimiento de genes o en su defecto genes huérfanos, para los cuales es necesario encontrar estrategias que ayuden a entender las dinámicas evolutivas de genes asociados al las variantes del sistema inmune no dependientes de inmunoglobulinas.

Esta tesis enfoca su estudio en el sistema inmune, que en sí mismo plantea asumir varios retos computacionales para tratar de responder a la siguiente pregunta: ¿cómo se define un gen del sistema inmune?, ya que como se expuso en la introducción general, debido a la amplia definición de lo que es la inmunidad, el conjunto de genes asociados es por tanto amplio y capaz de cumplir las funciones de mantener la integridad genética y biológica de los organismos, concepto bastante amplio que representa un desafío a nivel computacional al momento de establecer los límites claros que permitan diferenciar proteínas exclusivas del sistema inmune y proteínas, que aunque con una función importante, son accesorias a la función inmune del organismo. Es por tanto que proponemos como punto de partida en esta tesis el uso de bases de datos asociadas con la inmunidad característica de los tunicados como es el conjunto de genes asociados al SII. Uno de los problemas afrontados durante el desarrollo de esta tesis fue que las bases disponibles especializadas asociadas al sistema inmune, presentan un sesgo hacia las especies de humano y ratón debido a su importancia clínica. Como compensación decidimos complementar estas anotaciones con información con bases de datos asociadas al SII de insectos. Afortunadamente muchas de estas bases de datos están asociadas a bases de datos como el Ensembl que tiene un aspecto más amplio en cuanto que especies que se ven representadas y por tanto se pueden ampliar las inferencias que se proponen construir en este trabajo aunque se mantenga un sesgo hacia la información genómica y proteómica de vertebrados lo que implica que especies de invertebrados marinos como los tunicados se encuentren pobremente representadas.

Las bases de datos iniciales de las cuales se parte en este trabajo responden a lineamientos que permiten establecer límites en la clasificación de componentes del sistema inmune. Algunos de estos lineamientos se encuentra propuestos por la base de datos *IRIS* [59] en la cual las

proteínas contenidas tienen verificadas su asociación al sistema inmune según la literatura. Con base en estas proteínas, *IRIS* género un marco para definir un gen como perteneciente o no del SII basándose en seis principales parámetros:

- Filtrar si la función conocida o supuesta hace parte de la inmunidad innata o adaptativa.
- Establecer si participa en el desarrollo o maduración de componentes del sistema inmune
- Conocer si la proteínas es inducido por inmunomoduladores
- Saber si la proteína se expresa principalmente en tejidos inmunes
- Saber si participa en una vía inmune que resulta en la expresión de moléculas de defensa,
- Establecer si produce una proteína que interactúa directamente con agentes patógenos o sus productos.

Estos parámetros fueron útiles para escoger las principales bases de datos idóneas del Sistema inmune innato y a si poder establecer el set inicial de genes que permitan obtener un panorama general de los dominios y arquitecturas asociadas al sistema inmune.

1.1. Bases de Datos asociados al sistema inmune innato

Una de las bases ampliamente consultada para estudiar el SII es `InnateDB`[19], la cual es una base de datos de mamíferos que tiene como objetivo ser una base de conocimiento y plataforma de análisis de las redes moleculares complejas, rutas metabólicas e interacciones que ocurren entre las proteínas y moduladores propios del sistema inmune innato.
Las bases de datos que están asociadas a `InnateDB` deben cumplir cumplir con unos estándares mínimos de información requerida para reportar interacciones moleculares (minimum information required for reporting a molecular interaction experiment o MIMIx) y además de esto debe reportar la evidencia de cada interacción, el tejido o célula en donde la interacción fue reportada, el tipo de interacción y método de detección[19]. Debido a esta rigurosidad en la identificación de moléculas asociadas con la inmunidad innata y al ser uno de sus objetivos la construcción de una lista completa de genes del SII que se encuentren publicados que importa datos de múltiples bases de datos de genomas, de interacciones y de rutas metabólicas, escogimos esta base de datos como punto de partida en la definición de arquitecturas de dominios que representen un aproximación a lo que es un gen del sistema inmune innato pueda ser desde una visión de sus estructura. Entre las bases asociadas a `InnateDB` se encuentran:

- `Immport`: es una base de datos de inmunología, formada con el fin de colectar la mayor cantidad de información en humano relacionada al sistema inmune. En esta base de datos hay aproximadamente 6000 genes de humano, esta base de datos se generó a partir de búsquedas automatizadas en EntrezGene y ontologías usando palabras claves con algún tipo de relación a la inmunología, esta lista fue curada a mano por expertos acompañado de una confrontación con la literatura[11].

- `Septic Shock Group`: tiene como objetivo el entendimiento en detalle de la la inmunidad innata del cuerpo humano en defensa a diferentes enfermedades, enfocados principalmente en la respuesta de macrófagos a infecciones microbianas. Los genes que se encuentran en esta base de datos están acompañados de estudios de cambios significativos de expresión[**?**].

- `MAPK/NFKB Network`: es producto de la curación manual de las rutas de señalización p38, ERK, JNK , del módulo de regulación trascripcional NFKB y de datos provenientes de Transpath. [11]

- `Calvano et al., Nature 2005`: datos basados en datos de micro arreglos cuyo fin era encontrar cambios en los patrones de expresión génica en leucocitos sanguíneos sujetos a estímulos inflamatorios, en diferentes tiempos. Se encontraan 3.714 genes con cambios significativos[25].

- `Immunome Database`: los genes presentes en esta base de datos se obtuvieron a partir de artículos científicos y libros de texto sobre inmunología, la lista contiene solo los genes que están involucrado directamente en procesos inmunológicos. Se excluyeron los genes que están representados en todas las células, o que se encuentran incompletos, adjuntando solo los genes que participan en cascadas relacionadas con inmunidad. En total representan 844 genes.

Adicionalmente se utilizó la bases de datos de inmunidad innata de insectos `Insect Innate Immunity Database` (IIID) [20]. Aunque existen fuentes de diversas especies de insectos, del conjunto total de especies almacenadas en esta base de datos se tomaron aquellas que compartieran anotaciones tanto en `NCBI` como del `Ensembl` y que corresponden a las especies *Nasonia vitripennis*, *Apis mellifera*, *Drosophila melanogaster* y *Anopheles gambiae*,
Durante este estudio definimos las arquitecturas gold standard o arquitecturas canónicas del sistema inmune , como aquellas estructuras proteicas generalizadas que representa las estructuras proteicas más comunes de la transcripción de un gen en particular y que sirven como marco de referencia. Es por ello que debido a la rigurosidad que implica validar información experimentalmente y confrontarla con la literatura, la consideramos durante nuestro estudio como arquitecturas gold standard o arquitecturas canónicas a los genes provenientes de las bases de datos `InnateDB` y del `Insect Innate Immunity Database` (IIID).

La predicción de estas arquitecturas depende en gran medida del acceso que se tenga de anotaciones confiables como el Ensembl que posee un espectro amplio en cuanto a las especies representadas (incluyendo algunos tunicados) y las anotaciones están basadas en una pipeline de anotación robusta y ampliamente reconocida como se describe en la sección siguiente.

1.1.1. Pipeline de anotación del Ensembl

Entre los sistemas de anotación usados para resolver la anotación en los diferentes genomas de las especies *Ciona intestinalis* (Ciin), *Ciona savignyi* (Cisa), *Latimeria chalumnae* (Lach), humano y ratón usados como referencia en esta tesis, se encuentra la pipeline de anotación de Ensembl. En esta base de datos se confluyen multiples procesos que permiten identificar los elementos funcionales en una secuencia de DNA. Una de las características fundamentales de la anotación del Ensembl es que dicha anotación está soportada por la evidencia experimental característica y especialmente algo que la diferencia de otros métodos que se basan predicción ab initio es que está integrada con información filogenética de otras especies. La pipeline de ensembl se basa en la automatización de pasos computacionales que le permiten la integración de relaciones entre diversas entidades para la anotación que posterior permite la toma de decisiones por los curadores manuales, permitiendo de esta forma anotar miles de genes en paralelo, siendo la velocidad y la consistencia la clave de esta metodología. [92], [1]
Esta metodología se puede dividir en cuatro fases: la preparación del genoma para ser anotado, la construcción de modelo de proteinas codificantes, el filtrado y finalmente la presentación del conjunto de genes anotados. No todos los pasos del proceso son necesarios y dependen de la posición de la proteína evaluada dentro la filogenia construida para la especie, la calidad de ensamblaje del genoma, la importancia del organismo como modelo y la disponibilidad de los datos experimentales existentes para la especie de estudio[92], [1].

Preparación del genoma

En primer lugar se realiza el llamado de listas de contigs, scaffolds o cromosomas dependiendo del estado del ensamblaje de cada genoma. Para todos estos se cargan sus sistemas de coordenadas generando un sistema de referencia basado en etiquetas sobre coordenadas virtuales de la región mejor ensamblada que se conoce como toplevel [92], [1]. Subsecuentemente se da el enmascaramiento del genoma mediante el uso de los programas Repeatmasker basado en librerías de RepBase, Dust y Tanden repeat Finder. A su vez también se pueden construir modelos para genomas nuevos con la suite Repeatmodeler y estos modelos se comparan con las proteínas curadas de Uniprot y se eliminan las proteínas repetitivas. Subsecuentemente se hacen varios procesos de enmascaramiento tomando las bases de datos que maximicen el enmascaramiento[92], [1].
A continuación son producidos los modelos de genes por medio de *genscan* [24]. Estos modelos de genes son comparados por medio de Blast con las bases de datos de Uniprot y Unigene.

Entre otros componentes detectados como exones o CDs también se predicen islas CpG y sitios de inicio de transcripción entre otros [92], [1].

Posteriormente los datos ab initio son confrontados por alineamientos de cDNAs, proteínas y datos de RNAseq al genoma ensamblado y se priorizan los datos provenientes de la misma especie que cuenten con algún tipo de anotación con relación a secuencias predichas automáticamente en otras especies para así obtener como resultado final una base de datos tanto de secuencias alineadas como de un conjunto largo de proteínas codificantes candidatas. [92], [1]

Pipeline dirigida

Esta pipeline busca reducir el espacio de búsqueda para el generador de modelos de regiones codificantes usando el programa*GeneWise* [14]. por medio de dos procesos. En el primer proceso se corre con sitios de empalme con consenso y la otra con sitios de splicing. Este proceso se logra por la evaluación de proteínas propias de la especie y su alineamiento con el genoma. Para este proceso se usa un conjunto de proteínas de alta confianza obtenidas de Uniprot y Refseq, en donde se escogieron los números de acceso de las proteínas probadas en laboratorio (PE niveles uno y dos para Uniprot y NP y AP para refseq) de la especies en estudio. En una segunda pipeline (pipeline de similitud) se usa la misma estrategia solo que con proteínas de diferentes especies[92], [1].

Para casos experimentales de la misma especie se usa *pmatch*. Entre los parámetros de este programa se encuentra T14 que indica el número de aminoácidos consecutivos que deben dar match con el DNA genómico. En el caso de las proteínas que no son de la misma especie se usa BLAST directamente para poder identificarlos en el genoma. Se clasifican los hits de acuerdo a los niveles de Uniprot de existencia proteica o protein existence (PE) de Uniprot, las cuales esta clasificadas en diferentes niveles de acuerdo a la evidencia que se tenga (proteína, Trascrito, inferida por homología, Predicha o incierta), otras bases de datos sirvieron para la clasificación, de igual forma se uso la cercanía taxonómica con otras especies para priorizar las especies que están evolutivamente y así filtrar la búsqueda siendo este método eficiente cuando se tiene alta identidad entre secuencias a comparar en el proceso de anotación.

Para el conjunto de secuencias no alineadas por *pmatch* se usa una evaluación con *Exonerate* en donde el hit inicial se extiende 200 kb en ambas direcciones. Luego es usado el *algoritmo de empalme consciente* o *Splice-aware aligners* en ingles, generando modelos de regiones codificantes mediante el alineamiento de cDNAs y CDS que permite la identificación no solo de la región codificante, sino que a su vez permite identificar las regiones UTRs. Entre los múltiples modelos generados se escogen los modelos cuyo marco de lectura genere el menor número de codones de parada, dichos modelos se alinean localmente nuevamente, teniendo así un modelo final de transcrito[92], [1].

De forma paralela se usa una pipeline basada en lecturas de secuenciamiento de RNA seq,

las cuales se alinean contra el genoma usando BWA y en donde cada hit corresponde a un exón. Este conjunto de posibles exones son considerados proto-transcriptomas los cuales se vuelve a alinear con los RNAs de transcritos disponibles para buscar los sitios de corte entre exones e introns de nuevo por medio de *exonerate*. Durante este proceso se tiene en cuenta la cobertura y longitud de las lecturas, ya que entre más alta sea la cobertura es más probable encontrar los sitios de splicing de la totalidad de los introns. En caso contrario, si la cobertura de la lectura es muy baja la pipeline genera dos grupos de secuencias, la primera que la conforman los modelos transcripcionales y la segunda los no transcripcionales. En muchos casos para identificar las proteínas codificantes se usó un blast de identidad 80 % y una cobertura del 80 %, enriquecidos con información transcripcional de varios tejidos para así tomar la transcripción del mismo gen en varios tejidos y lograr tener un consenso de componentes en el transcriptoma final de la especie [92], [1].

Pipeline de ortología

Cuando las pipelines dirigida y de similitud no logran generar todos los modelos de genes, se procede a usar la pipeline de ortología, la cual está basada en un blast recíproco entre proteínas de otras especies cercanas y de esta forma se complementan los modelos de anotación truncados. Como resultado se tiene los datos de ortología interconectados con todas las especies. [92], [1]

Pipelines de proyección

Las pipelines dirigidas y de similaridad no son útiles en genomas fragmentados de baja cobertura donde faltan regiones genómicas por ensamblar, no está claro el orden del fragmento en el contexto global del genoma original y mucho menos la ubicación de genes o quizas muchos genes están representados parcialmente o se encuentran fragmentados en varios scaffolds. Para solucionar estos inconvenientes se desarrolló la metodología de alineamiento global del genoma de referencia, el cual normalmente es el genoma del humano (whole genome alignment (WGA)) por medio de *BLASTZ* para alineamientos locales que luego son puestos en el orden correcto consecutivamente por medio de la herramienta *axtTools*. Esta herramienta también tiene en cuenta evitar sobreposiciones de fragmentos, es decir que cada una de las bases de datos a comparar no tenga más de una posición en el genoma. También se usan las anotaciones de genes para resolver el orden correcto de los genes en los scaffolds objetivos de comparación [92], [1].

Identificación de las regiones no traducidas o UTRs

Debido a que los programas de generación de transcritos dependen de cDNA, no es posible calcular las UTR directamente, así que una vez identificados los genes codificantes son usados estos para la adicione la regiones 5 y 3 UTRs basados en información proveniente de

experimentos de tipo CAGE y ditags los cuales permiten identificar las posiciones de inicio y final de la transcripción para asi compararlos con los modelos preestablecidos de cDNAs.[92], [1]

La pipeline de sensibilidad como su nombre lo indica tiene como objetivo la construcción de posibles modelos de transcripción en donde se hace énfasis en la sensibilidad, es decir en poder identificar los verdaderos positivos, y luego es seguido por un proceso de filtrado donde se concentra en la especificidad, es decir en la capacidad de identificar los verdaderos negativos, posteriormente son seleccionados los modelos con mayor confianza. [92], [1].

Al alinear transcritos de especies lejanas con un genoma de referencia por medio de la pipeline de anotación similaridad se generan errores en el programa *Genewise* sobre los modelos de splicing generados. Este inconveniente es solucionado mediante el filtro de transcrito consenso o *TranscriptConsensus* en inglés. Este consenso es basado en mínimo seis modelos ya sean derivados de cDNAs, modelos de RNA-seq y de EST de la misma especie en donde son comparados los límites de los exones e intrones y se pondera la calidad de los modelos producidos de novo por la pipeline de similaridad conforme a la longitud y puntuación que estos generan [92], [1].

LayerAnnotation

En esta fase del proceso se filtran los datos provenientes de las pipelines dirigidas y de los datos de RNA-seq. En esta etapa se genera un conjunto de datos jerárquicos en donde los más altos son los datos provenientes de la misma especie seguido por datos de especies filogenéticamente cercanas y por último se evalúan los modelos que hayan dado un buen puntaje por *TranscriptConsensus*. Los modelos que no queden entre estos conjuntos no son seleccionados y los modelos seleccionados son pasados por la pipeline *GeneBuilder* cuyo objetivo es eliminar los modelos redundantes y dejar los transcritos alternativos para cada gen [92],[1].

Pseudogenes

Para todos los modelos producidos por *GeneBuilder* se evalúan si existen pseudogenes. Estos se identifican cuando existen copias con un solo exón que indican que pudo procesarse de retrotranscripción de un mRNA de rivado de un pseudogen procesado o copias de pseudogenes truncados que poseen codones de parada e intrones retenidos o copias con muchas repeticiones que puede indicar la creación de un intrón que no estaba en la copia original y en otros casos de intrones muy cortos[92],[1].

La incorporación de aplicaciones externas de anotacion manual como *Havana* produce un set de genes fusionados entre los genes producidos por la pipeline del Ensembl y las propias producciones de *Havana*. De igual forma los modelos modelos de GENECODE y Refseq se utilizan para garantizar constantemente las actualizaciones de los modelos de Ensembl

conocidos como Consensus Coding Sequence (CCDS) y que son considerados los más estables y confiables de la anotación[92], [1].

Biomart y Ensembl

Debido al auge de las nuevas tecnologías de secuenciamiento, se ha evidenciado un incremento de los datos biológicos en los repositorios públicos, debido a esto se hizo necesario software que facilitaran el acceso e integración de estos datos para los análisis bioinformáticos, en este punto surgió, BiomaRt el cual es un proyecto de creación de software y de datos a la comunidad científica, está basado en el código abierto y permite la recuperación de grandes cantidades de datos de manera uniforme sin necesidad de hacer consultas por SQL, [37]. Entre las bases de datos de BiomaRt esta Ensembl, dándoles acceso a los usuarios de BiomaRt acceso directo a los datos de Ensembl permitiendo hacer consultas en línea o entre otras opciones se encuentra usando el paquete de `Bioconductor` el cual es un paquete de `R` de `biomart`, el cual permite acceder a los datos y descargarlos desde el software de R. [38]

1.2. Identificacion de dominios del sistema inmune

Una vez conocido el sistema de anotación implementado para los genes de las especies de referencia utilizadas en este trabajo que conforman `InnateDB` y `Insect Innate Immunity Database` (IIID) nos vimos enfrentados a un nuevo problema computacional para poder construir las arquitecturas de dominios de genes del SII de referencia. Gran parte de la predicción de las estructuras canónicas de proteínas depende de la asignación correcta de los dominios que las componen, para este propósito, diferentes consorcios han construido bases de datos basadas en el uso de modelos HMM de dominios proteicos. Cada una de estás bases es independiente y sobre un gen es posible tener asociados todos los dominios detectados por cada uno de los predictores de cada base o en su contrario solo uno o en muchas ocasiones dos. Entre éstas la más conocida es `Interpro` que en una sola base de datos integran múltiples bases de datos de HMM como lo son: PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, GENE3D y PANTHER.
Dichos modelos al ser combinados en un solo sitio de búsqueda, evitan la redundancia de información, que es aplicada a nuevas estructuras proteicas con el fin de tener una aproximación a su caracterización funcional. Debido a la aproximación probabilística de los HMM que es más robusta en comparación con las búsquedas de similitud de los alineamientos locales de Blast, partimos de los HMM provenientes de estas base de datos ya que nos garantizan la calidad de los modelos, sino que intrínsecamente hay una homogeneidad entre los números de acceso de las diferentes bases de datos.
Entre las propuestas que existen para la detección de homología, la más popular es el alineamiento local de BLAST, sin embargo se ha encontrado que esta estrategia tiene debilidades a la hora de establecer homólogos remotos o genes que tiendan a duplicar dominios o a

realizar reordenamiento de exones como ocurre con genes del sistema inmune. Este proceso natural de crear nuevas combinaciones de exones mediante recombinación intrónica se denomina exón shufle [61]. Este proceso está directamente relacionado a genomas más grandes y menos compactos los cuales presentan un aumento en la longitud de sus regiones intronicas siendo más frecuente el exón shufle en estos genomas, funcionando como un motor evolutivo, por ejemplo, en la radiación acelerada de los metazoos facilitando la aparición de proteínas multimodales de superficie celular [**?**] como los receptores trasmembranales asociados con la inmunidad. Debido a que el barajamiento de exones le da diversidad a la proteína permitiendo el reordenamiento de sus exones conservando su función, que presenta dificultades al momento de ser evaluada por homología directa.
Es por ello que durante esta tesis se propone el uso de modelos probabilísticos tipo HMM para poder inferir de mejor manera los homólogos remotos por medio de la anotación de dominios basados en arquitecturas proteicas de referencia obtenidas del `InnateDB` y `Insect Innate Immunity Database` (IIID). La anotación por medio de los dominios que componen una proteína es bastante fiable, pero debido a las singularidades del sistema inmune, como estar sometido constantemente a fuerzas de selección propias de cada especie, a las dinámicas evolutivas de los dominios que componen esas arquitecturas, hace del proceso de detección de posibles conjuntos de arquitecturas de dominios un proceso a ser evaluadas con sumo cuidado, ya que pueden existir proteínas que contengan el mismo tipo de dominio pero el orden y el número de repeticiones varían drásticamente de una a otra, generando convergencias evolutivas o innovaciones génicas cuya función debe evaluarse posteriormente experimentalmente.

1.2.1. Modelos Ocultos de Markov implementados en bases de datos de dominios de proteínas

Una de las ventajas del uso de proteínas para la anotación de genes, es que sus estructuras poseen los elementos principales para realizar su función y pueden representar una ventaja al momento de evaluar homologías remotas o estructuras de genes complejos ya que sólo están compuestas por las regiones traducidas de los ORF organizadas en secuencias de dominios. Por ejemplo una de las bases de datos mal ampliamente utilizada es el Pfam en el cual se ha generado alineamientos bastante robustos sobre los cuales se han construido HMM capaces de encontrar homologías remotas.
Sin embargo, a pesar de los problemas mencionados anteriormente frente a la anotación de genes en genomas fragmentados, (la posibilidad de tener fragmentado un gen en contigs diferentes), los HMM permiten identificar pequeñas secuencias homólogas. Los HMM son modelos estadísticos probabilísticos que pretende modelar parámetros desconocidos ocultos en un conjunto de estados para posteriormente emplearlos en el reconocimiento de los patrones buscados. De esta manera, el alineamiento múltiple es construido a partir de un conjunto de proteínas o motivos homólogos que sirven para entrenar el modelo y que busca tener la

probabilidad de poder obtener cualquier secuencia de símbolos a partir del mismo.[39]. Entonces, el alineamiento múltiple - de nucleótidos o aminoácidos - puede ser representado como un modelo probabilístico denominado perfil HMM, donde cada posición tiene una probabilidad asociada a cada uno de los estados, y a su vez las posiciones están interconectadas por probabilidades de transición entre ellos. Siendo posible calcular la probabilidad de cambios en la arquitectura proteica tales como delecion o inserción de dominios representadas en saltos al siguiente nodo o en cambios en la distribución de la probabilidad de los símbolos de salida.
Los HMMs de dominios de proteínas se encuentran reportados en las siguientes bases de datos:

- **Pfam**: es una colección de modelos ocultos de Markov (HMMs) de familias de proteínas construidos a partir de alineamientos múltiples en donde también se encuentran representados clanes proteicos y tienen una proteína de UniProtKB de referencia [44]

- **TIGRFAMs**: contiene alineamientos múltiples de secuencias curadas, HMM de proteínas, e información diseñada para soportar anotaciones automatizadas. Además, provee el número de accesión, nombre de la proteína y tipo de modelo [50],[51].

- **SUPERFAMILY**: es una base de datos de anotación estructural y funcional de proteínas. Dicha anotación está basada en la colección de HMMs, que representan los dominios proteicos, generando grupos de superfamilias basados en relaciones evolutivas. Esta base de datos está construida sobre más de 2.478 genomas secuenciados[119].

- **PIRSF**: presenta un agrupamiento de secuencias no redundantes de UniProtKB basados en relaciones evolutivas, su clasificación está basada en dos aspectos, la homología (que provengan de un ancestro común) y la homeomorfología (similitud de la secuencia completa de un dominio compartido). [120]

- **PANTHER** (Protein ANalysis THrough Evolutionary Relationships): es un sistema de clasificación de proteínas que se clasificaron acorde a su su relaciones evolutivas por familias, subfamilias (con la misma función), función molecular y contexto biológico. Es decir, se clasifican proteinas que pertenezcan a una misma ruta metabólica o que se encuentre en procesos biológicos similares [111].

- **CATH-Gene3D**: esta base de datos describe familias de proteínas y arquitectura de dominios a lo largo de los genomas, generados a partir del algoritmo de agrupamiento de Markov, seguido de agrupamiento de vecindad multiple (multi-linkage clustering) basados en la identidad de la secuencia [90].

- **HMAP**: los HMM son creados manualmente por curadores que identifican familias o subfamilias de proteinas conservadas [113].

- **UNIPROT**: es una base de datos que almacena información sobre la función de las proteínas, arquitectura de dominios, interacciones, vías biológicas, variantes genéticas, entre otras. Cada entrada de Uniprot no solo contiene secuencias proteicas con anotaciones curadas, sino que la mayoría de la información es obtenida de literatura científica lo que garantiza una validación experimental [40].

Aunque se encuentren múltiples bases de datos con HMM de proteínas, la forma en las que se generaron y se curaron hacen imposible el cruce entre ellas ya que son estandarizadas con parámetros propios e independientes entre las bases, sumado a que por ejemplo el Uniprot que entre sus objetivos tiene integrar todas estas bases de datos aún posee inconvenientes en cruzar correctamente los diferentes modelos, debido a estas razones se recomienda evaluar cada una de ellas de forma independiente como lo hemos hecho en este trabajo.
Teniendo en cuenta los problemas relacionados con la anotación *de novo* de genes y la robustez de los HMMs para identificar dominios, consideramos que éstos se vuelven una unidad fundamental para ser buscados y fundamentales para la anotación de genes del sistema inmune. Debido a que la estructura canónica básica de ellos se caracterizan por estar constituidas por un conjunto de dominios altamente conservados a lo largo de las diferentes ramas de la evolución pero ordenados y combinados de forma diferente [88].. Estos dominios representan la mayor fuente de diversificación del sistema inmune ya que las diferentes combinaciones entre ellos generan un universo casi ilimitado de proteínas a partir de un número limitado de dominios. Dichas combinaciones permiten mantener patrones de estructuras canónicas de receptores del sistema inmune entre especies, pero con variaciones en el número y tipo de dominios que constituyen dichas proteínas, de tal manera que permitan al organismo moldear una respuesta inmune que responda a las necesidades que el ambiente le propone.
Así, ya que partiendo de un conjunto limitado de dominios se logra tener un universo casi ilimitado de proteínas debido a sus posibles combinaciones mediadas por *exonshuffling* [89], es necesario el desarrollo de una pipeline que automatice la identificación *de novo* de dominios del sistema inmune en genomas de especies no modelo, como los tunicados.

1.2.2. Metodos de detección de dominios tipo HMM: HMMER

La estrategia comunmente usada para la identificación de secuencias por similitud es BLAST que se basa en una heurística de alineamientos locales basadas en el algoritmo de Smith-Waterman. Este método ampliamente utilizado, ha demostrado tener problemas al momento de detección de homologías remotas [118]. Además, los HMM son sensibles para búsquedas de similitud remota debido inferencia probabilística que usa en la comparación de secuencias, además ayuda al establecimiento de homología en secuencias altamente divergentes. Una herramienta como HMMER es ampliamente utilizada ya que realiza la búsqueda de dominios posición por posición y la puntuación de los gaps son basados en un perfil de consulta y calcula la señal de homología basada en el algoritmo de HMM de *avance-retroceso* el cual

no sólo calcula la mejor puntuación de alineamiento sino que suma el soporte de todos los posibles alineamientos[43] [118].
La aproximación HMMER es una combinación de varios algoritmos, el algoritmo striped vector-parallelized alignment, el algoritmo (SIMD single instruction of multiples data), un instructor de vectores llamado SSE, sumado a algoritmos de aceleración heurístico y un método conocido como 'sparse rescaling' que permite procesar el algoritmo HMM de *avance-retroceso* que permite ejecutar instrucciones de multiplicación y de sumatorias de probabilidades escaladas sin un subdesbordamiento aritmético o numerical underflow, es decir, donde los resultados del un cálculo es un valor absoluto menor al que el computador puede almacenar en la CPU. En dicho algoritmo los puntajes de avance, que son puntajes de verosimilitud se basan en el logaritmo de Cuota (log-odds likelihood scores) sobre un alineamiento incierto [43]. Estos puntajes de avance son útiles en la detección de homólogos remotos debido a que a menudo pueden haber múltiples formas de alinear una secuencia problema distante pero que a su vez se encuentra relacionada evolutivamente (secuencia query) con una secuencia objetivo (secuencia target), ya que hace la sumatoria de todos los posibles alineamientos, lo que implica que cada alineamiento contribuye al puntaje que permite indicar similaridad[43].

1.3. Metodología

1.3.1. Construcción de un conjunto de dominios de referencia gold standard

Con miras a construir una posible solución a dos problemas asociados con la anotación *de novo* de genes asociados al sistema inmune que mencionamos como 1) Heterogeneidad de las estructuras génicas de genes asociados al SI y 2) ¿Cómo lograr predecir arquitecturas *ab initio* capaces de cumplir con las definiciones de *IRIS*[59] y que sean comparables con las estructuras canónicas reportadas en las bases de datos del sistema inmune innato?, planteamos una metodología para construir un sistema de dominios de referencia o *golden standard* (referidos de ahora en adelante como estructuras canónicas) y que fueron obtenidos a partir de los dominios de proteínas de genes anclados en las bases de datos: `InnateDB` [18] e `Insect Innate Immunity Database` (IIID) [20]. Estas bases de datos fueron escogidas, debido a que permiten acceder a genes y proteínas cuya función en la inmunidad innata se encuentra experimentalmente validada, constituyéndose en una fuente de datos donde la función inmune está claramente definida como se explicó en la sección anterior. Durante este estudio nos inclinamos por el uso de la base de datos `InnateDB` debido a que en sí misma esta base de datos contiene más bases de datos como: `Immport`, `Immunogenetic related information source (IRIS)`, `Septic Shock Group`, `MAPK/NFKB Network`, `Calvano et al. Nature 2005` e `Immunome Database`, proporcionando un conjunto robusto de proteínas de referencia idóneas para ser el punto de partida de esta tesis.

Recuperación de dominios

Para obtener las secuencias de genes, proteínas e información general asociada a dichas bases de datos, se descargaron los números de accesión en formato tabular directamente del sitio `InnateDB`: `http://www.innatedb.com/redirect.do?go=resourcesGeneLists`. Para poder recuperar las secuencias y las anotaciones de los dominios que componen las proteínas del sistema inmune relacionadas con `InnateDB` filtrando la información contenida Gene ID del `Ensembl`.
De igual forma se usó la base de datos `IIID` para obtener la información asociada a genes y proteínas del sistema inmune innato en algunas especies de insectos. Del conjunto total de especies almacenadas en esta base de datos se tomaron aquellas que compartieran anotaciones tanto en `NCBI` como en `Ensembl`, las especies contenidas en esta base de datos que cumplen esta condición fueron *Nasonia vitripennis*, *Apis mellifera*, *Drosophila melanogaster*, *Anopheles gambiae* y *Acyrthosiphon pisum*. De esta base de datos filtraron los números de acceso de `NCBI` y se cruzaron los ID con los números de acceso de `Ensembl Metazoa` (`http://metazoa.ensembl.org/index.html`) a través de un script en `Perl`.
Con los código de acceso del `Ensembl` (para humano y ratón provenientes de `InnateDB`) (Dada la composición de la base de datos del `InnateDB` la mayor cantidad de número de accesos recuperados correspondieron al humano, versión genómica (GRCh38) y al ratón (GRCm38)) y `Ensembl Metazoa` (para las proteínas de insecto provenientes de `IIID`), se procedió por medio del script en `R` usando el paquete `biomart` [38] a la descarga de las secuencias proteicas, números de acceso únicos del gen, transcritos, proteínas y las respectivas anotaciones según diferentes sistemas de anotación de los dominios proteicos asociados a dichas proteínas de inmunidad innata con sus respectivas coordenadas de inicio y final. Este procedimiento se realizó sobre Ensemblv.86.

Obtención de los HMM específicos para el Sistema inmune

Con el fin de construir un repositorio de HMM específicos del SI por cada base de datos previamente mencionada, que permita focalizar los recursos de cómputo en la caracterización de genes asociados al SI en las diferentes especies problema, se generó la metodología expuesta en la figura **1-1** la cual parte de la lista única de números de acceso de los HMM asociados al SI por cada una de las bases de datos anotadas en el `Interpro` `Pfam`, `Gene3D`, `PIRSF`, `Hamap`, `Superfamily` y `TIGRFAM` (Figura **1-1**) , se generó el script en shell hmmfetch.sh
Este script invoca la herramienta `hmmfetch` que toma números de acceso de los HMM y extrae esos números de acceso del archivo plano que contiene la totalidad de HMM proveniente de una base de datos por ejemplo como el Pfam, generando así un archivo plano con los HMM de interés,
`hmmfetch` fue ejecutado con los siguientes parámetros:

- `hmmfetch -o -` [Archivo de salida] [Base de datos con la totalidad de los HMM]

El primer argumento es como va a ser el archivo de salida, el segundo argumento es la base de datos donde se encuentra la totalidad de los dominios y el tercer argumento es el archivo plano con números de acceso de los dominios asociados al sistema inmune.
Debido a la ya mencionada complejidad en la conformación de arquitecturas proteicas que permitan el equilibrio entre conservar las propiedades del receptor y la adquisición de nuevas arquitecturas que le permitan una plasticidad a su medio ambiente se plantea establecer, es en las variaciones de este equilibrio que surgen ramas de la inmunes que son rastreables transversalmente a lo largo de la evolución y estructuras especie especificas, que como su nombre lo indica son únicas de la inmunidad de cada organismo ausentes en especies cercanas

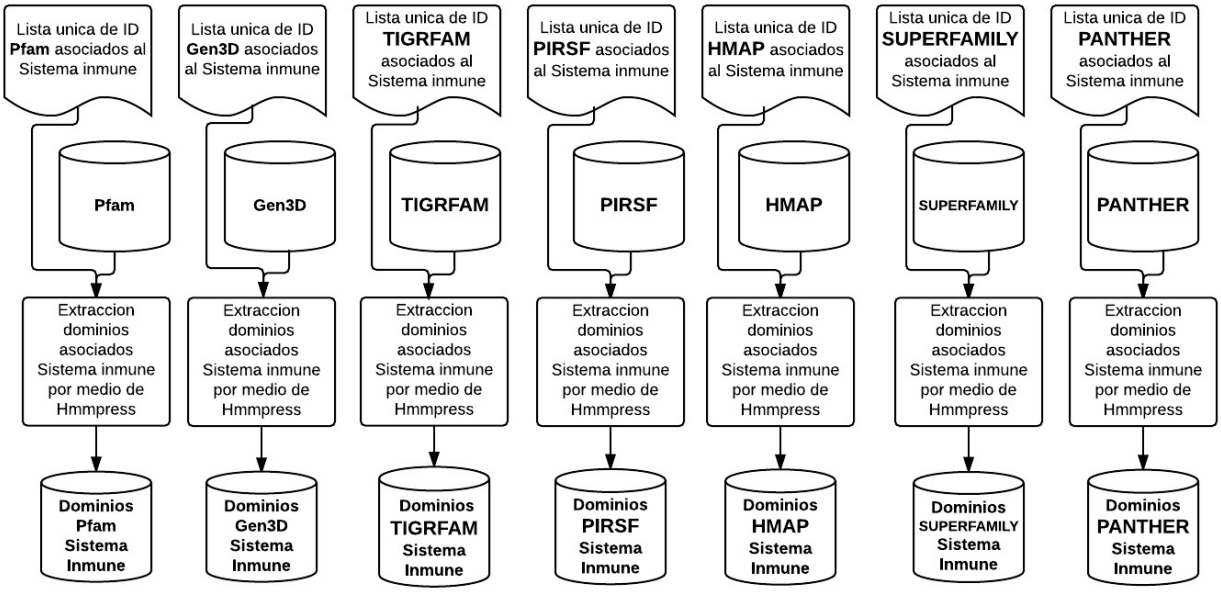

Figura 1-1: Mediante este flujo de trabajo se obtuvo la información contenida en la base de datos de Uniprot, entre esta información está los números de acceso de los HMM asociados al sistema inmune de las siguientes bases de datos:SUPERFAMILY,PIRSF,CATH,PANTHER,TIGRFAMs, HMAP,PFAM y el correspondiente número de acceso en la base de datos de Interpro que permite el cruce de información entre ellas

Con las arquitecturas colectadas anteriormente (en humano, ratón e insectos con una relación comprobada con la inmunidad) se definió el conjunto de arquitecturas canónicas (*gold standard*) del sistema inmune innato. Estas arquitecturas se componen de dominios de HMM procedentes de las bases de datos asociadas a interpro como: `SUPERFAMILY`[119], `PIRSF`[120], `Gene3D`[90], `PANTHER`[111], `TIGRFAMs`[51], `HMAP` [113] y `PFAM`[50]. Sin embargo ya que cada uno de estos sistemas de anotación dependen de HMM adaptados para la búsqueda de los criterios que dominan cada base de datos entonces en nuestra aproximación la distribución

de arquitecturas obtenidas por las bases de datos se define entonces a partir de la distribución de los dominios que han sido anotados de manera independiente por cada una de estos sistemas de anotación.
A partir de los modelos HMM recuperados por cada base de datos se procedió porsteriormente a realizar una búsqueda de dominios en proteínas de genes no anotados por el Ensembl utilizando la suite HMMer. Dependiendo del sistema de annotación por cada base se utilizaron los filtros reportados en cada uno de los esquemas propios de anotación.

1.3.2. Diseño de estrategias para detectar estructuras de dominios del gold standard en cordados

Sumado a las diferencias existentes en los criterios de anotación de los dominios en las bases de datos, es necesario abordar el problema intrínseco que tienen las arquitecturas génicas de genes del SI. Para resolver este problema se utilizaron tres aproximaciones las cuales analizadas en conjunto podrian solucionar parte de los problemas que se presentan en la detección de homologías de proteinas asociadas al SI debido no solo a la variación en el número de dominios y sino también al correspondiente orden. Estas tres estrategias se llamaran en su orden Estrategía de **O**rden de arquitecturas, Estrategía de **D**esorden de arquitecturas y Estrategía de **B**lasp. Se consideró en esta aproximación que las duplicaciones estrictamente consecutivas de dominios fuesen resumidas en una lista de dominios que fue ordenada por la coordenada de inicio pero que carece de repeticiones consecutivas. Para esto se diseño un filtro que se denomina *Función de reducción*. Con este tipo de aproximaciones se pueden hacer búsquedas más felxibles de las arquitecturas canónicas evitando las dificultades que generan las ganancias o perdidas de dominios en proteínas para las cuales se buscan homólogos remotos.

Estrategía de Orden de arquitecturas

Se definió la lista de arquitecturas canónicas de dominios *gold standard* para cada una de los sistemas de anotación HMM correspondiente. Dichos dominios predichos poseen diferente número de acceso del HMM del cual provienen y se encuentran ordenados por sus coordenadas genómicas de inicio.

Estrategía de Desorden de arquitecturas

Es sabido que muchos receptores del SI tienen variaciones en su arquitectura que aunque posean cambios en el orden de dominios esto no implica la perdida de su función; dichas variaciones están dadas por combinación, rearreglos, copias y perdidas de dominios, los cuales quedarían excluidos del análisis por orden. Es por esto que se decidió adoptar una estrategia de desorden de arquitecturas para afrontar este problema. En este caso cuando se realiza la comparación de composición de dominios en desorden de dominios *gold standard* y

desorden entre las dominios de las especies con anotación del Ensembl y con otro sistema de anotación. En este caso el orden entre los dominios no es determinante para poder asignar una proteína como candidata putativa al SII.

Estrategía de Blastp:

Como método de soporte adicional se aplicó la estrategia de `blastp` reportada en [63]. Esta estrategía se diseñó para buscar homologías entre secuencias de proteinas usando la implementación clásica de homologia en Blast. Como secuencias queries se utilizaron las secuencias de proteínas a partir de las cuales se construyó el *gold standard* y como secuencias blanco se usaron las proteínas de genomas anotados y no anotados por Ensembl. Para cumplir con este objetivo se ejecutó el programa con los siguientes parámetros:

- ***blastall -p blastp -d DB -i QUERY -f 9 -F m S -M BLOSUM45 -e 100 -b 10000 -v 10000 -m 8***
- blastall ***-p*** *Program Name* o nombre del programa que en este caso fue blastall
- ***tblastn -d*** *Data base*, la base de datos contra la que se va a comparar los dominios, en este caso el genoma en nucleótidos
- ***-i*** *Query File* secuencia problema o query que va a ser comparado contra la base de datos, en este caso los dominios en aminoácidos,
- ***-f*** *Threshold for extending hits* Número máximo de hits: *9*,
- ***-F m S*** Enmascaramiento leve
- ***-M*** *Matrix* matriz de sustitución de bloques de aminoácidos: *BLOSUM45*,
- ***-e*** *Expectation value (E)* Valor de e-value esperado en este caso se establecio como *1e-5*
- ***-b*** Número de mejores Hits en la region que se pretende evaluar , se escogió el *10000*
- ***-m*** *alignment view options* tipo de archivo de salida para visualizar de los resultados, se escogió el *8*.

Los resultados de esta búsqueda por blast fueron filtrados por medio del script en `Perl` filtro.pl cual tubo en cuenta los siguientes parámetros como criterio de inclusión de proteínas no anotadas que puedan estar asociadas a la arquitecturas canónicas.

- Un valor $E \leq 0{,}001$
- Valores de cobertura con respecto a la secuencia query $\geq 60\,\%$
- Identidad $\geq 30\,\%$

.

1.3.3. Obtención de dominios candidatos al sistema inmune en genomas de cordados con anotación del Ensembl

Al ser los tunicados un punto intermedio en el árbol evolutivo del grupo de los cordados, se analizaron los cordados más basales presentes en `Ensembl` (representados como input en la Figura **1-2**). Es importante aclarar que en estas especies aun existen muchas proteínas sin anotación funcional. Entonces, como punto de partida para identificar las proteínas y dominios asociados al SII se usó en primer lugar un script en `R`, utilizando `biomaRt` para recuperar todas las proteínas y las anotaciones de dominios con su respectivo inicio y final existentes en el Ensembl para las anotaciones genómicas de las especies de tunicados *Ciona intestinalis*, *Ciona savignyi* y de otros cordados inferiores como *Petromyzon marinus*, *Latimeria chalumnae* y *Danio rerio*. Estás especies se identificarán de ahora en adelante como Ciin, Cisa, Pema, Lach y Dare. En la Figura **1-2** se observa que el mismo procedimiento fue utilizado sobre las especies de referencia utilizadas para construir las arquitecturas canónicas derivadas de las bases `InnateDB` e `Insect Innate Immunity Database` (IIID).

Cruce de arquitecturas canónicas asociadas al sistema inmune

Una vez recuperadas las arquitecturas de dominios para las cinco especies de cordados que tienen anotación en el Ensembl, entonces por medio del script en `Perl` Comparacion_Info.pl, de forma automatizada se usaron las diferentes estrategias previamente explicadas, (**O**rden de candidatos, **D**esorden y **B**lastp) definidas desde la información de las arquitecturas canonicas *gold standard*. Se realizó un proceso de evaluación de comparación de estas diferentes aproximaciones obteniendo los candidatos putativos que fueran identificados clasificados por las tres estrategias (ODB) o que solo fueran identificadas como arquitecturas canónicas mediante las estrategias de orden y desorden (OD).
Es importante resaltar que los outputs de la figura **1-2** funcionan como los input de los esquemas resumen de las 3 estrategias **1-3** **1-4** y **1-5**.

1.3.4. Búsqueda de genes relacionados al sistema inmune en las especies de cordados sin anotación del Ensembl: los tunicados Oikopleura dioica y Botryllus schlosseri y el cefalocordado Branchiostoma floridae

Este grupo de organismos corresponde a las especies *B. floridae*, *B. schlosseri* y *O. disocia* las cuales se identificaran como Bsc, Oidi y Brfl respectivamente en este trabajo. Las anotaciones a nivel de proteínas de las anteriores especies fueron recuperadas de: `JGI Genome Portal` (`http://genome.jgi.doe.gov/Brafl1/Brafl1.download.ftp.html`) para *B. floridae*, `Oikoarrays` (`http://oikoarrays.biology.uiowa.edu/Oiko/Downloads.html`) para *O. dioica* y del sitio de descargas de la base de datos `Aniseed` (`http://www.aniseed.cnrs.fr/`

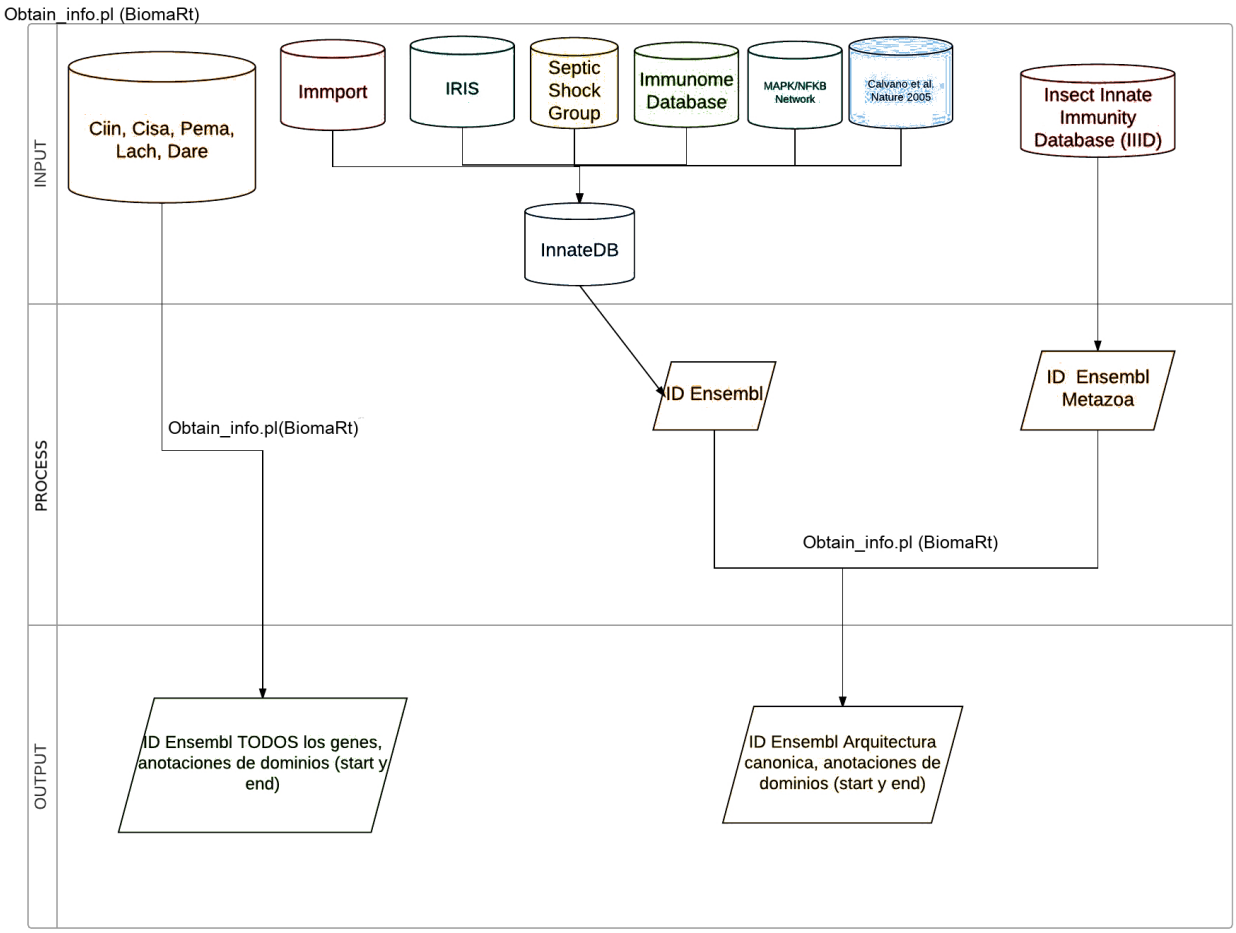

Figura 1-2: Mediante este flujo de trabajo se obtuvo las secuencias de proteínas, anotaciones de dominios con su respectivo inicio y final en las proteinas asociadas al sistema inmune según su composición de dominios para humano, raton e insectos y así como para todas las secuencias proteicas de las especies Ciin, Cisa, Pema, Lach y Dare

.

`aniseed/download/download_data`) para el caso de la especie *B. schlosseri*. El procedimiento de identificación de candidatos putativos se realizó a partir de los modelos HMM recuperados de los dominios de proteínas que conforman el sistema *gold standard*. Con estos se realizó una búsqueda automatizada de dominios sobre la lista de proteínas de estas especies utilizando el programa HMMer. Posteriormente sobre las secuencias candidatas se aplicaron de nuevos las estrategias de **O**rden de candidatos, **D**esorden y **B**lastp para obtener una lista de proteínas candidatas a ser asociadas con el SII.

Los datos obtenidos fueron filtrados por medio del script en `Perl` filtro.pl expuesto anteriormente, como se evidencia en la gráfica figura **1-6**.

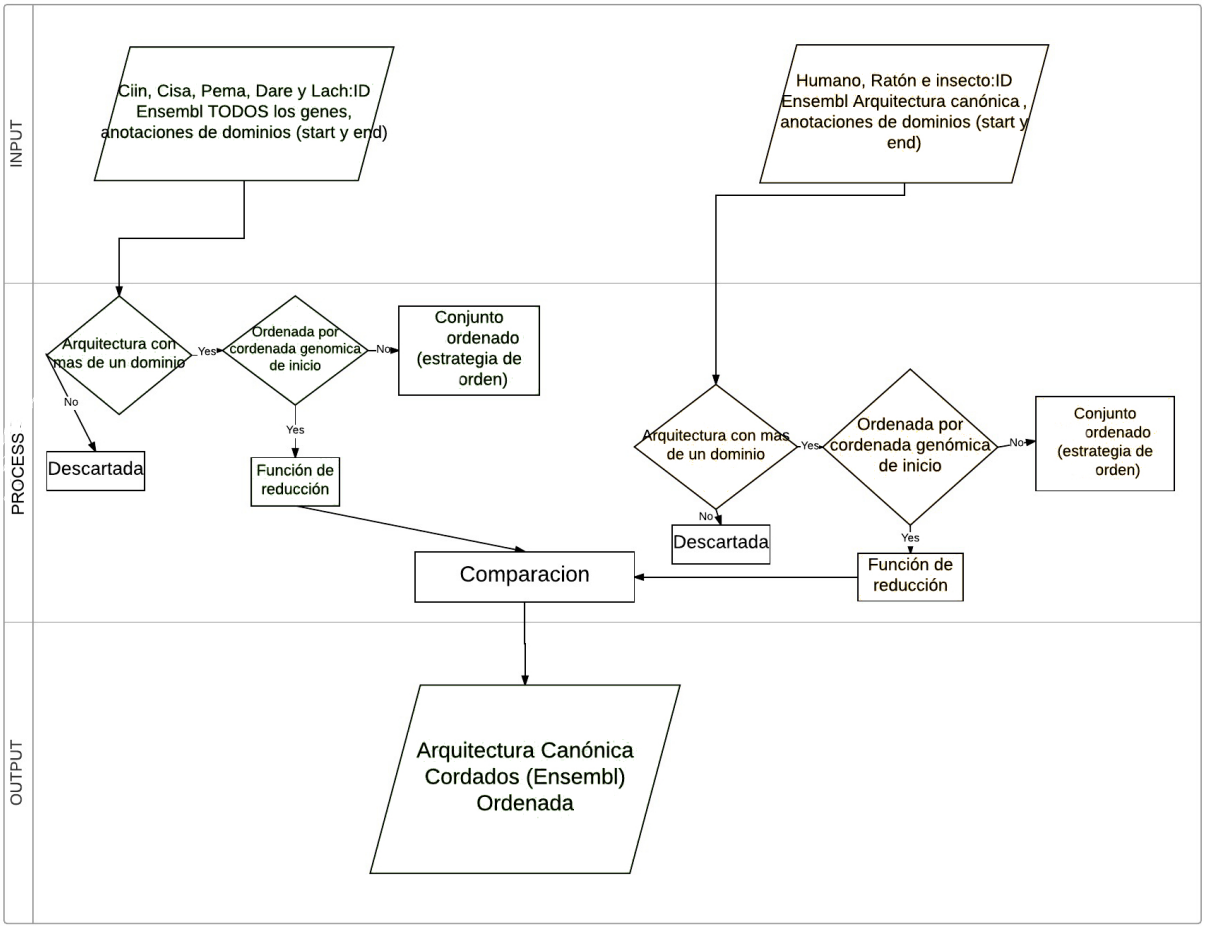

Figura 1-3: Diagrama que explica el proceso de comparación entre la arquitectura canónica asociadas a la inmunidad innata en humano, ratón e insectos con las proteínas de cordados basales de Ensembl utilizando la estrategia de **O**rden

1.3.5. Definición de los dominios asociados a los módulos de Señalización, Efector y Reconocimiento del sistema inmune

Para definir los dominios que pertenecen a las diferentes módulos asociados del sistema inmune se usaron los datos provenientes del estudio realizados por Zárate Potes en el 2014, Dicho estudio hizo una exhaustiva búsqueda bibliográfica de proteínas asociadas al sistema inmune y las clasificó de forma manual en los tres módulos Señalización, Efector y Reconocimiento. A partir de dichas listas se descargaron las proteínas relacionadas tanto de Uniprot y y NBCI RefSeq. Posteriormente se identificaron las arquitecturas proteicas basados en información de Uniprot Swissprot y Pfam. Fue allí donde se estableció la lista final de dominios asociados a cada uno de los módulos que fueron filtradas sobre la lista de candidatos putativos obtenidos para los dominios asociados al SII de las especies de estudio [122], el procedimiento se

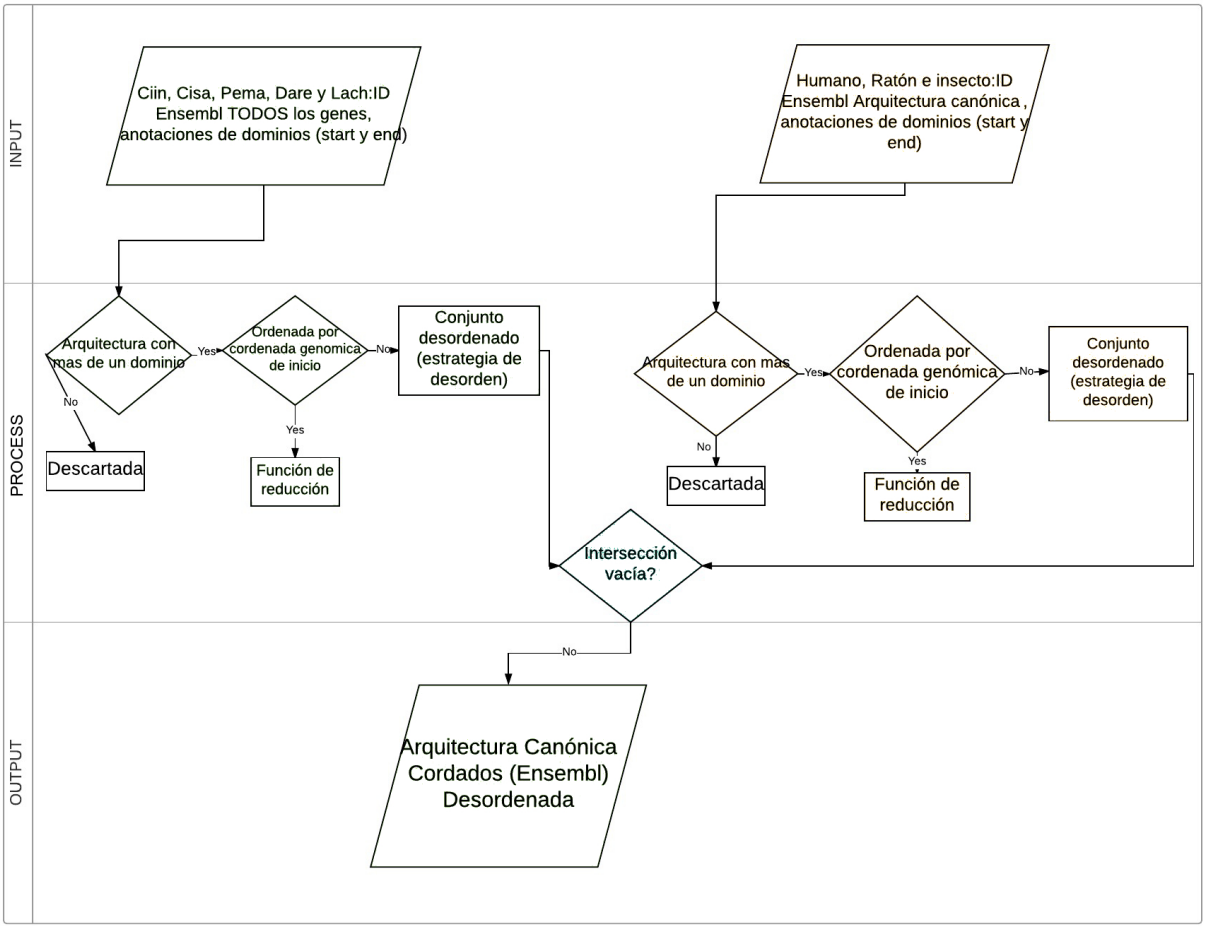

Figura 1-4: Diagrama que explica el proceso de comparación entre la arquitectura canónica asociadas a la inmunidad innata en humano, ratón e insectos con las proteínas de cordados basales de Ensembl utilizando la estrategia de **O**rden

observa en la figura **1-7**.

1.4. Resultados

1.4.1. Estado de ensamblaje de los Genomas del Grupo de Estudio

El estado actual del ensamblaje de cada una de las especies trabajadas en este estudio se observa en la gráfica **1-8**. Se evaluó en un intervalo de 30,000 Mb cuantos fragmentos genomicos se ven representados. Para los genomas de las especies Bosc, Lach, Ciin se observan valores con longitudes de fragmentos ensamblados mayores a 30,000 y en los casos como Brfl y Cisa hay diferentes tamaños en los fragmentos. Se logra observar que la especie *D. vexillum* poseen aun un ensamble de genoma muy limitado en el cual la mayoria de los fragmentos

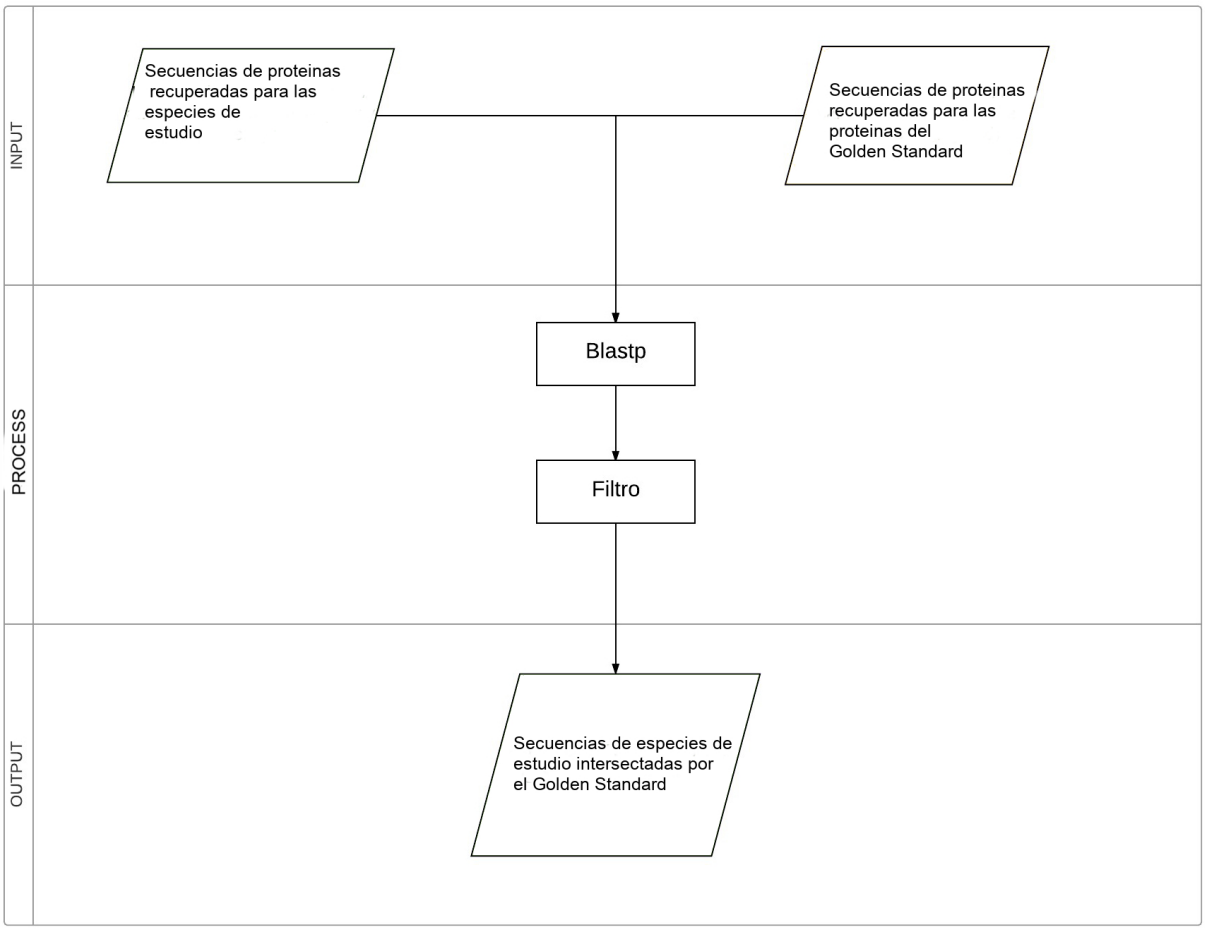

Figura 1-5: Metodología empleada para la evaluar la estrategia complementaria de `blastp` entre las proteínas canónicas asociadas a la inmunidad innata en humano, ratón e insectos con el fin de establecer homologías remotas.

son menores a 30,000 bp. Es importante resaltar que esta especie recibe especial atención en este trabajo ya que su genoma no se encuentra aún anotado y será el objeto de estudio en el capitulo 2.

1.4.2. Dominios asociados al Sistema inmune en cada base de datos

La cantidad de dominios asociado a cada una de las bases de datos de modelos de marcov se observa en la Tabla **1-1**, en donde se puede observar que en la primera columna está relacionada la base de datos donde fueron obtenidos los modelos asociados a los dominios, la segunda columna muestra el total de HMM asociados a Dominios por cada una de las bases de datos y por último la cantidad de ese total que se encontraron en estructuras canónicas del SII de las estrucucturas Gold standard.

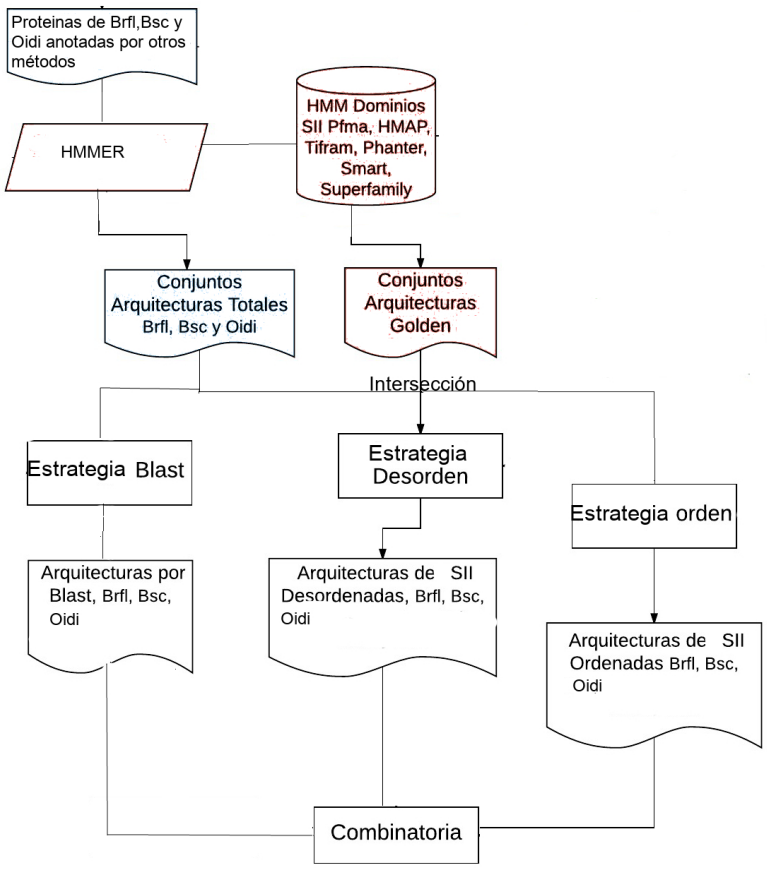

Figura 1-6: Flujo de trabajo para la obtención de las arquitecturas para las especies *Brfl*, *Bosc* y *Oidi* a partir de la predicción de dominios asociados al SII obtenidos previamente por medio de las arquitecturas Golden Standard en humano, ratón e insectos

En esta tabla se observa que Pfam tiene un total del 256902 dominios, siendo la base de datos con mayor número de estos modelos, segudos por CATH con 175987 y SUPERFAMILY con 173919. En cuanto a los dominios asociados al SII, se destaca que la base de datos Phanter [112] se reportaron 10124 dominios seguida por Pfam 3615, por el contrario se observa que la base de datos que menos dominios reporto fue PIRSF con 603.

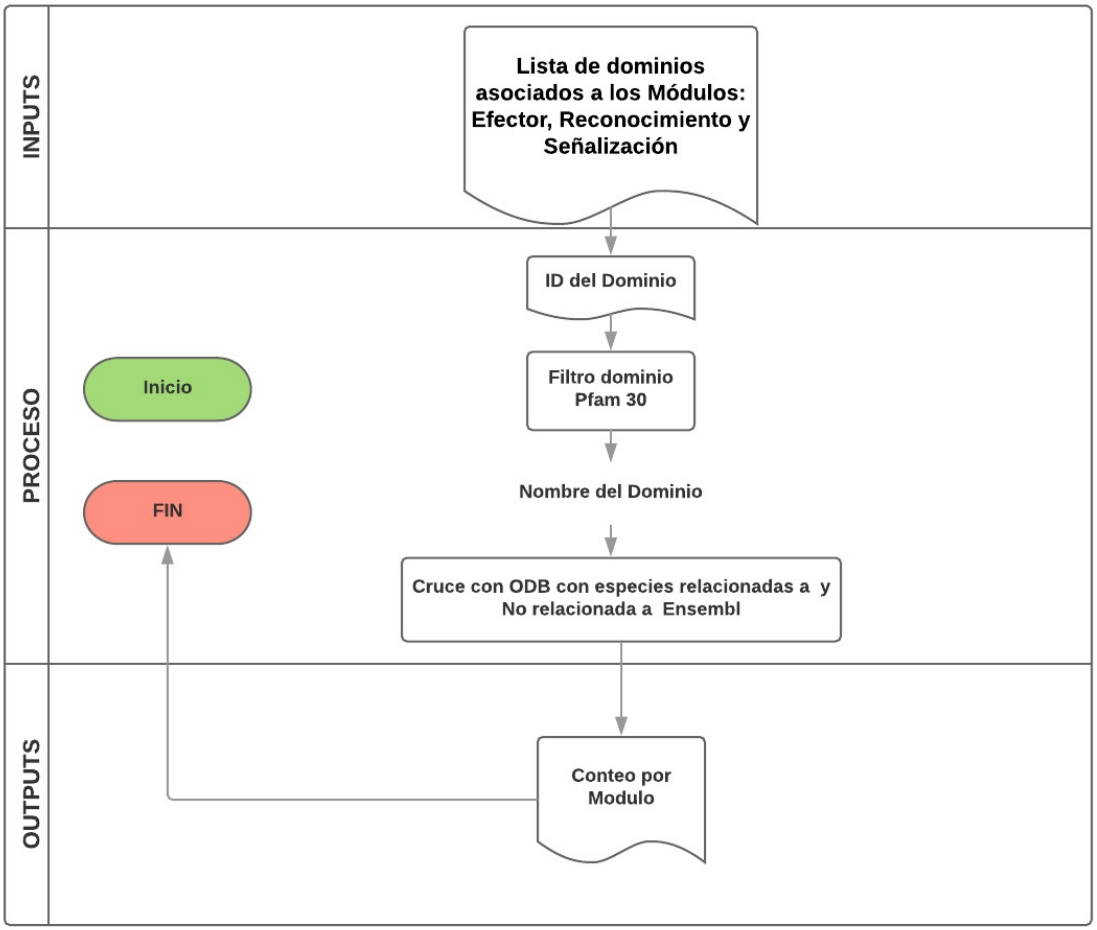

Figura 1-7: Diagrama que explica el proceso de como se generaron los conteos por modulo asociado al sistema inmune **O**rden

Tabla 1-1: Número de Dominios asociados al Sistema inmune innato en las arquitecturas canonicas o golden standard después de usar hmmfetch, evaluado en cada una de las bases de datos .

Base de Datos de Dominios	Total de Dominios	Total Dominios golden del SII
SUPERFAMILY	173919	897
PIRSF	11287	603
CATH	175987	709
PANTHER	60851	10124
TIGRFAMs	22474	344
PFAM	256902	3615

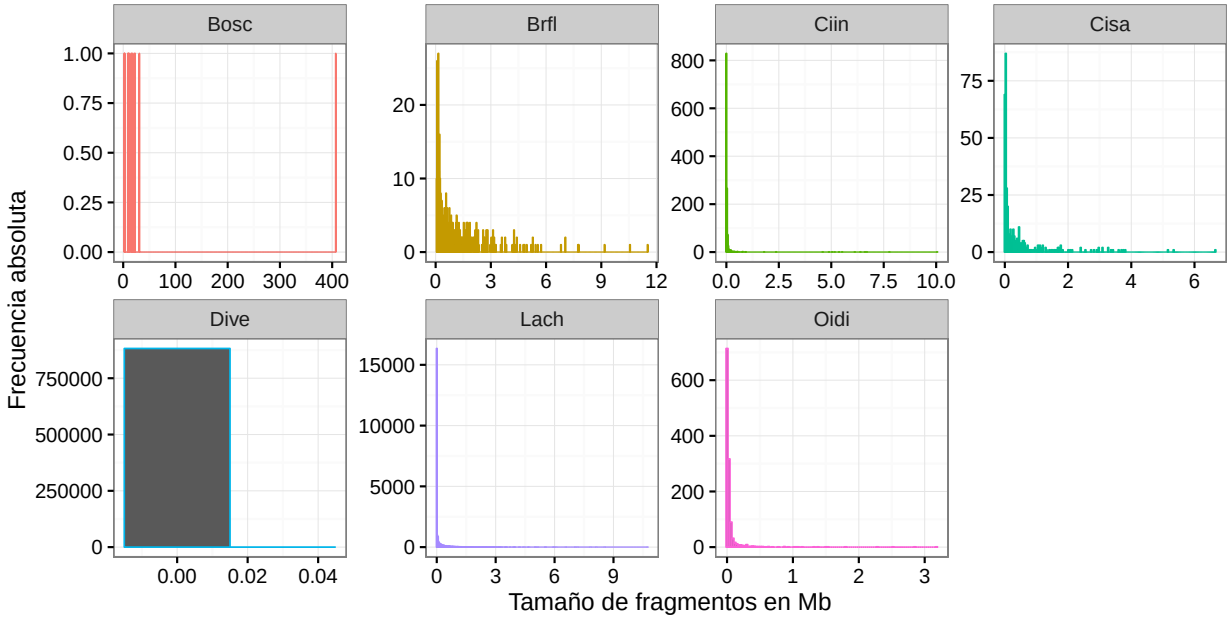

Figura 1-8: En esta grafica se muestra la distribución de los tamaños de las diferentes unidades de ensamblaje en las que se encuentran los diferentes genomas utilizados en este trabajo

En la gráfica **1-9** se observa cuántos de los HMM de cada base de datos de Uniprot se encuentra representado en las arquitecturas canónicas del SII, se aclara que no cuenta el número repeticiones de cada HMM sino su presencia o ausencia dentro de la totalidad de las arquitecturas canónicas, se observa de igual forma que la base de datos que más se destaca es Phanter seguida de Pfam y Prosite. También se observa que el resto de bases de datos se encuentran por los mismos valores entre 600 y 800 dominios reportados.

En la Tabla **1-2** se observa la frecuencia de dominios que se pueden identificar luego de haber aplicado las estrategias de anotación **O**rden, **D**esorden y **B**last (**ODB**), se destaca que la base de datos con mayor número de dominios reportados despues de ser aplicada las diferentes estrategias es Pfam con 1195, seguida por SUPERFAMILY con 448 y CATH con 313 mientras que en las bases de datos Smart, Phanter, Prosite y PIRSF están completamente ausentes.

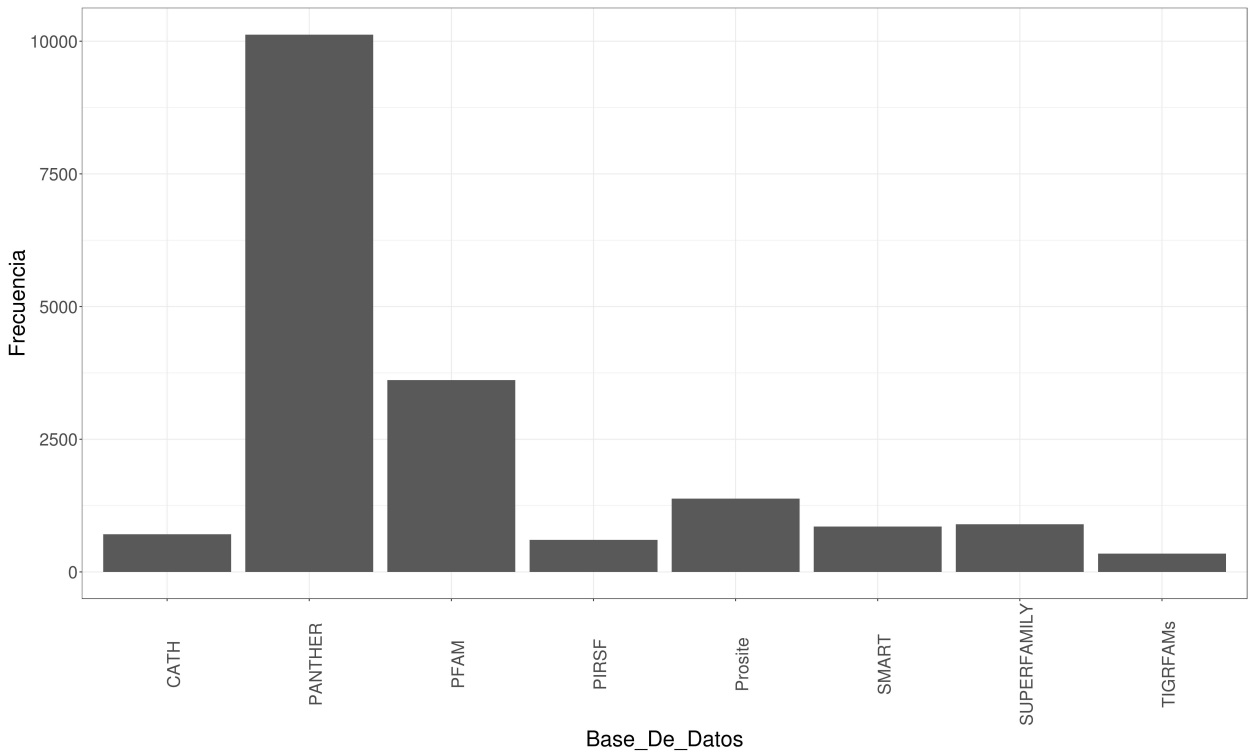

Figura 1-9: Muestra la cantidad de dominios asociados al sistema inmune encontrados en las arquitecturas Golden, relacionados con cada base de datos de donde se extrageron

Tabla 1-2: Número de dominios usados en la anotacion por medio las estrategias ODB por base de datos

Base de Datos de Dominios	Total Dominios golden del SII	Total Dominios Anotación
SUPERFAMILY	897	448
PIRSF	603	0
CATH	709	313
PANTHER	10124	0
TIGRFAMs	344	36
PFAM	1382	1195
Prosite	1382	0
Smart	857	0

En la figura **1-10** se observa la frecuencia de dominios que se pueden identificar luego de haber aplicado las estrategias de anotación **O**rden, **D**esorden y **B**last (**ODB**), es interesante ver

como solo las bases de datos CATH, PFAM y SUPERFAMILY son las únicas que presentan dominios que pueden ser hallados en las anotación de nuevas arquitecturas mientras que en las bases de datos Smart, Phanter, Prosite y PIRSF están completamente ausentes.

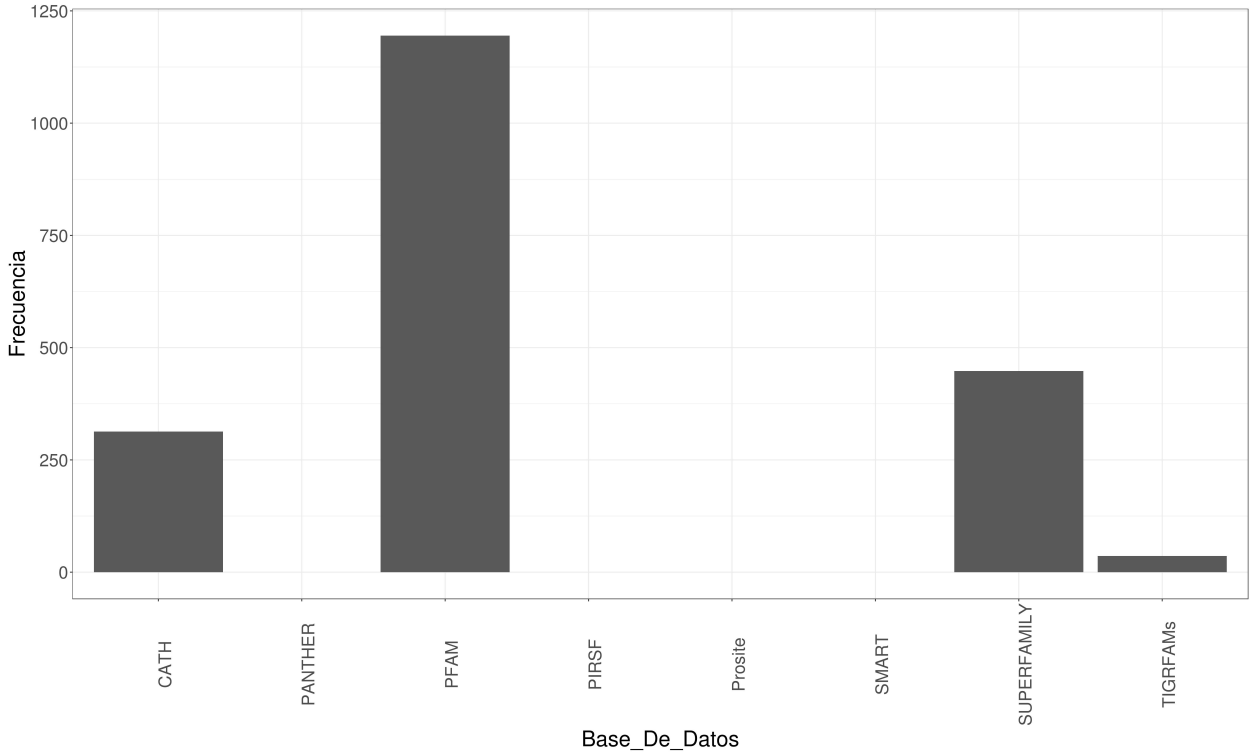

Figura 1-10: Muestra la frecuencia de dominios asociados al sistema inmune encontrados en las arquitecturas Golden despues de aplicar las diferentes estrategias de anotacion

1.4.3. Comparación de las estrategias de anotación Orden,Desorden y Blast (ODB)

Con el fin de visualizar cuales combinación de estrategias o estrategia única fueron las más efectivas a la hora de proponer candidatos asociados al sistema inmune basados en el uso de las arquitecturas canónicas o golden standard y para visualizar el impacto que tienen las diferentes estrategias sobre el proceso de anotación de candidatos putativos al SII se realizaron los conteos por cada especie y por cada estrategia. En las siguientes secciones se presenta el porcentaje de anotaciones por cada estrategia y por cada especie. En los anexos se resume en la Figura **5-1** los resultados obtenidos en los que se presentan las diferentes frecuencias de cada una de las estrategias evaluada por cada uno de las especies que pudieron ser asociadas con estructuras canónicas *gold standard*.

Comparación de las estrategias de anotación Orden,Desorden y Blast por especie

- *Acyrthosiphon pisum*: *Acpi* Para poder evaluar qué estrategias, ya sea de forma individual o en combinatorias, fueron las más efectivas para la anotaciones de proteínas en las diferentes especies de estudio (*Ciin*, *Cisa*, *Lach*, *Pema*, *Bosc*, *Oidi* y *Brfl*) por medio de la comparación de sus arquitecturas con las arquitecturas golden standard o canónicas provenientes de *Acpi*, se generó el grafico **1-11**, el cual, con miras a normalizar los datos presenta las diferentes combinatoria de estrategias que permitieron anotar proteínas del SII en las especies de interés, en esta gráfica se puede observar que la estrategia **OD** logró identificar arquitecturas, aunque en bajo porcentaje, produciendo anotaciones en las especies: *Bosc*, *Cisa*, *Lach*, *Oidi* con 39.9 %, 31.8 %, 47.1 %, 47.4 % respectivamente. De igual forma se evidencio que la estrategia **D** logró identificar arquitecturas, aunque en bajo número, produciendo anotaciones en las especies: *Brfl* , *Oidi*, *Pema* con 26.5 %, 46.1 %, 40.6 % respectivamente.

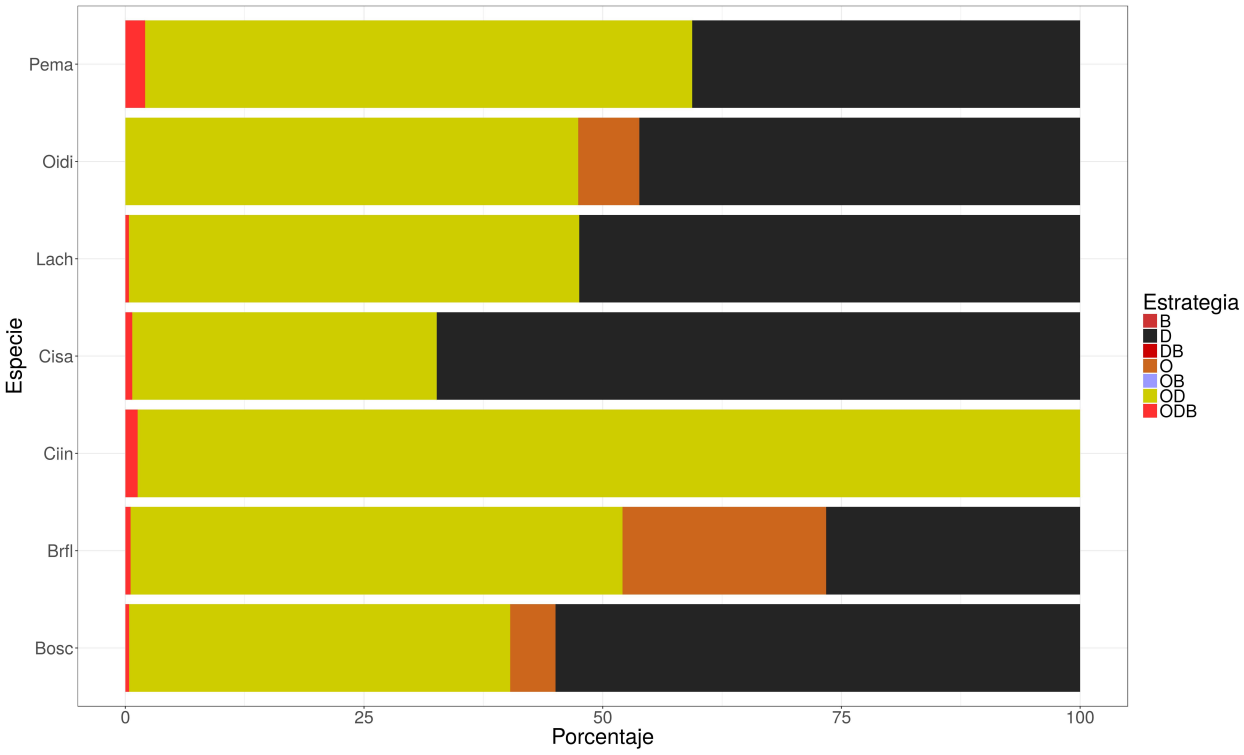

Figura 1-11: Porcentaje de las diferentes estrategias de anotación: **O**rden, **D**esorden y **B**last (**ODB**) *Acyrthosiphon pisum* con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

En cuanto a la estrategia **O** logró identificar arquitecturas, aunque en bajo proporción, produjo anotaciones en las especies: *Bosc* con 4.7 %, *Oidi* con 6.4 % y un poco mas sobresaliente en *Brfl* con 21.3 % . aunque en las especies especies: *Ciin*, *Cisa*, *Lach* y *Pema* no logró predecir ninguna arquitectura.

Al momento de comparar las estructuras canónicas provenientes de *Acpi* se evidencio que mediante la estrategias **OB**, **DB** No fue una estrategia exitosa para anotar genes del SII en las especies: *Bosc*, *Brfl*, *Ciin*, *Cisa*, *Lach*, *Oidi*, *Pema*, de igual forma la estrategia **ODB** no logró predecir ningún tipo de arquitectura en la especie *Oidi*

se encontró que la estrategia **ODB** logró aunque en bajo número, identificar arquitecturas en la especie *Pema* con un 2 %, pero sin duda el dato más sobresaliente se encontro con la estrategia **OD** ya que se predijo mediante esta estrategia el 98.7 % en *Ciin*.

- *Anopheles gambiae*: *Anga*: Al normalizar los datos con porcentajes para la especie *Anga*, como se observa en la gráfica en la figura **1-12**, que la estrategia combinada **ODB** fue efectiva, aunque en bajos porcentajes, logrando anotar genes en las especies: *Bosc* 24.7 %, *Oidi* 24.5 %, *Cisa* 14.8 % *Brfl* 11.5 %, pero siendo menos efectiva en *Ciin* 5 %, *Pema* 2.2 %. por otro lado la estrategia **OD** mostró una tendencia similar generando anotaciones las especies: *Bosc*, *Brfl*, , *Cisa*, *Oidi* con 28 %, 20 %, 33.3 %, 11 % respectivamente, siendo *Ciin* el caso más sobresaliente con un porcentaje del 72.5 %.

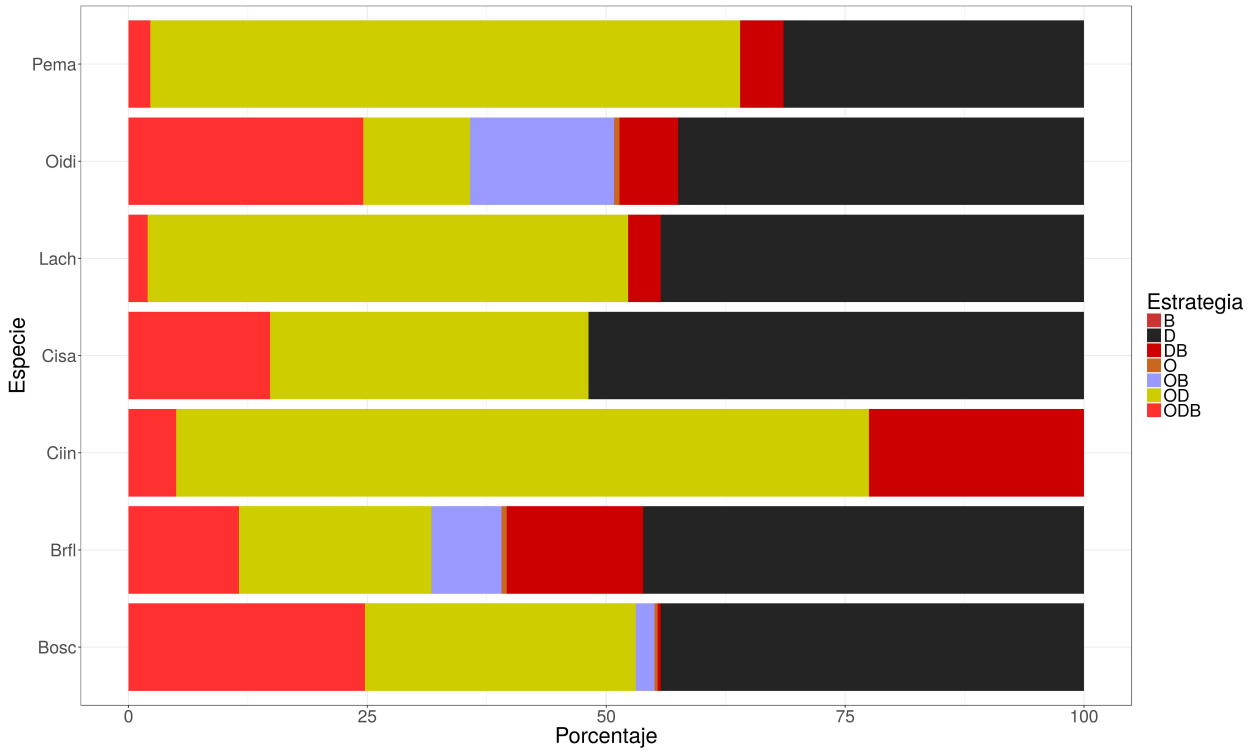

Figura 1-12: Porcentaje de las diferentes estrategias de anotación: **O**rden, **D**esorden y **B**last (**ODB**) *Anopheles gambiae* con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

Por otro lado, la estrategia **OB** logró identificar arquitecturas, aunque en bajo número, produjo anotaciones en las especies: *Brfl*, *Oidi* con 7.3 %, 15 % respectivamente, pero

totalmente ausente en las especies *Ciin*, *Cisa*, *Lach*, *Pema*, al compararse con otras estrategias combinadas como **DB**, se ven similitudes, ya que al igual que con textbfOB logra identificar arquitecturas pero en baja proporción, siendo anotadas por medio de la estrategia **DB** en las especies: *Brfl* con 14.2 %, *Ciin* 22.5 %, *Lach* con 3.42 %, *Oidi* 6 % y *Pema* con un 4.5 % aunque, es de resaltar que esta estrategia no logró anotar ninguna arquitectura en la especie *Cisa*

Al comparar las estrategias inviduales se observa que **D** logró identificar arquitecturas, aunque en bajo número en las especies: *Bosc*, *Brfl*, *Lach*, *Oidi*, *Pema* con 44.3 %, 46.1 %, 44.3 %, 42.5 %, 31.4 % respectivamente, por el contrario, la estrategia **O** no logró anotar ningún gen del SII en las especies: *Ciin*, *Cisa*, *Lach*, *Pema*

- *Drosophila melanogaster*: *Drme*

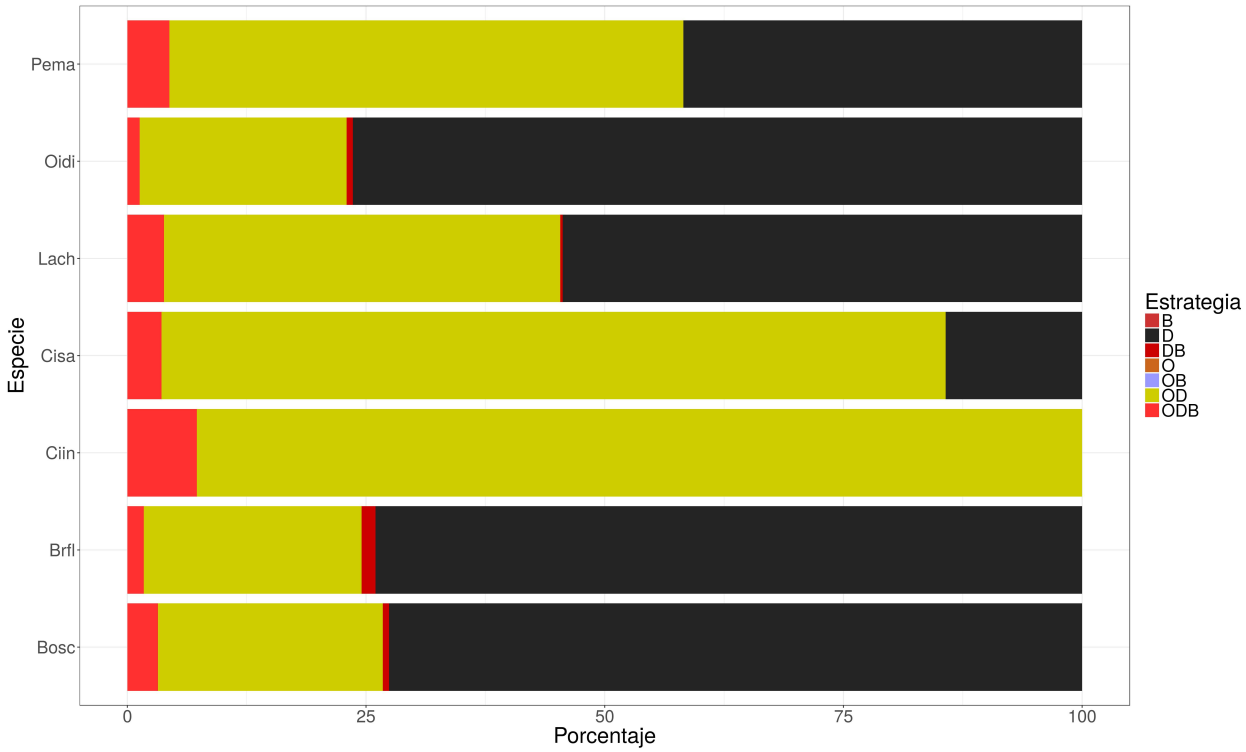

Figura 1-13: Porcentaje de las diferentes estrategias de anotación:**ODB** *Drosophila melanogaster* con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

Al evaluar las estructuras canónicas de *Drme* en la grafica **1-13**, Se observó que las estrategias con algún tipo de combinación con la estrategia a **D**, tuvieron cierto nivel de éxito al anotar tunicados, como es el caso de la estrategia **ODB** logró identificar arquitecturas, aunque en forma muy reducida, en todas las especies, por el contrario, la estrategia **OD** tubo resultados sobresalientes como *Ciin* 92.7 % y *Cisa* con 82.1 % y

en menor medida en especies como *Bosc* (23.55 %), *Brfl* (22.8 %), *Lach* (41.4 %) y *Oidi* con 4, 2 %, sin la estrategia de **DB** no dio ningun resultado apra las especies *Ciin*, *Cisa* y *Pema*

Como dato sobresaliente encontramos que la estrategia **D** fue significativa para la especie *Oidi* al lograr anotar el 76.3 % de las arquitecturas halladas, de igual forma aunque en menor proporción esta estrategia logró identificar arquitecturas en las especies: *Cisa* 14.2 %, *Pema* con 41.7 %, en contraposicion se observo que la estrategia basada exclusivamente en Orden **O** no logró anotar ninguna arquitectura en las especies *Bosc*, *Brfl*, *Ciin*, *Cisa*, *Lach*, *Oidi* y *Pema*, de igual forma ocurrio con la estrategia **OB** que no logró anotar arquitecturas en las especies *Bosc*, *Brfl*, *Ciin*, *Cisa*, *Lach*, *Oidi* y *Pema*

- *Apis mellifera*: *Apme*

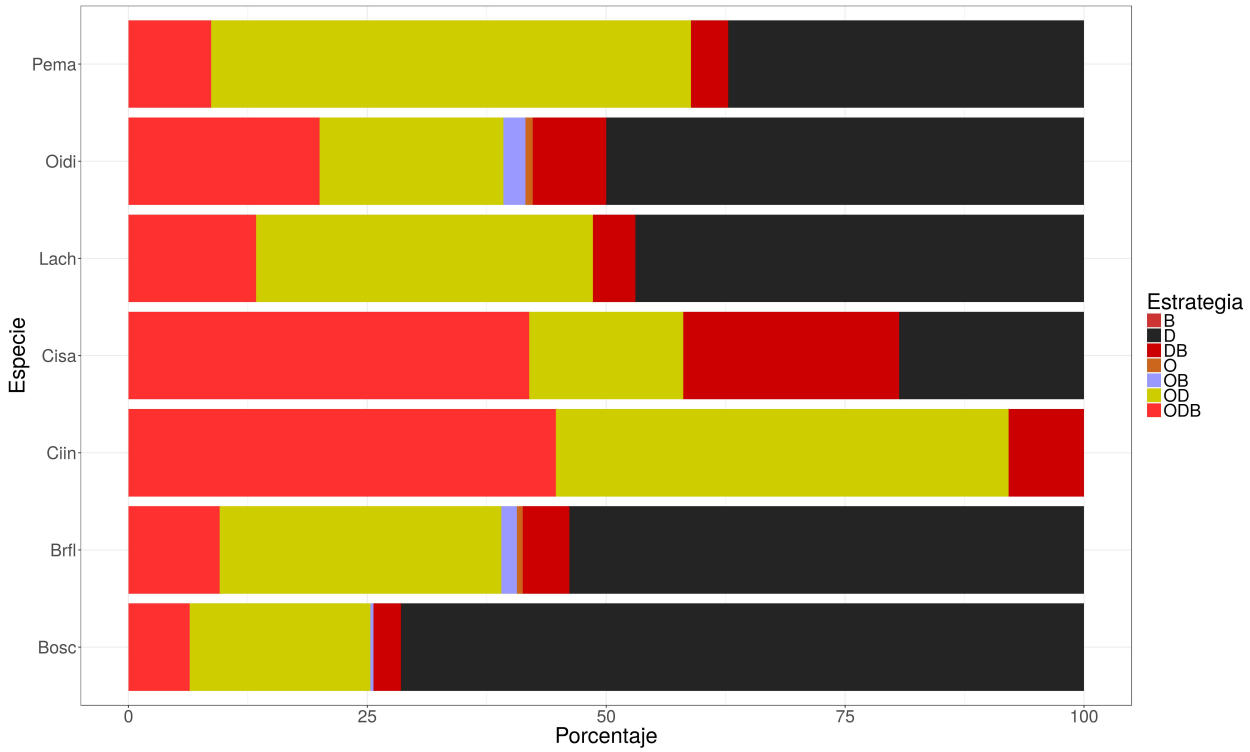

Figura 1-14: Porcentaje de las diferentes estrategias de anotación: **O**rden, **D**esorden y **B**last (**ODB**) *Apis mellifera* con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

De igual manera, se evaluaron las estrategias individuales como textbfD y **O**, donde se evidencio que en la primera fue bastante exitosa en especies como *Lach* y *Pema* pero en menor medida en *Cisa*. mientras tanto, la segunda estrategia no produjo ninguna anotación en las especies *Bosc*, *Ciin*, *Cisa*, *Lach* y *Pema*.

- *Nasonia vitripennis*: *Navi*

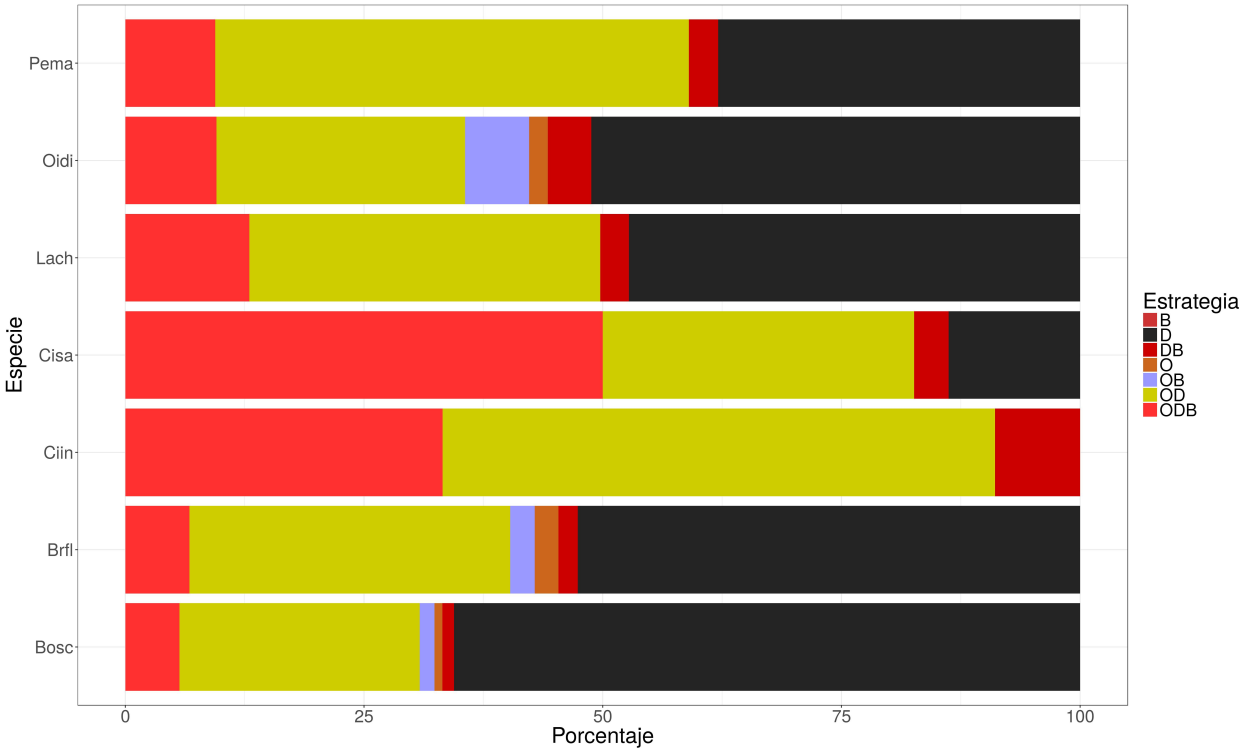

Figura 1-15: Porcentaje de las diferentes estrategias de anotación:**ODB** *Nasonia vitripennis* con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

Al evaluar las estructuras canónicas de *Navi* en la gráfica **1-15** se observa que las estrategias basadas en la estrategia **D** mostraron ser las más efectivas, es el caso de la estrategia **DB** textsf*Brfl* (2 %), *Ciin* (8.9 %), *Cisa* (3.6 %), *Oidi* (4.5 %), *Lach* y *Pema* ambos con 3 %. pero fue la estrategia **D**, que mejores resultados dio, encontrándose en : *Lach*, *Pema* con 13.7 %, 47.2 % y 37.9 %, respectivamente. de igual forma se observa que otra combinación de estrategias como **ODB** logró identificar arquitecturas en las especies: *Bosc* (5.6 %), *Brfl* (6.7 %), *Ciin* (33.2 %), *Lach*(13 %), *Oidi* (9.5 %) y *Pema* (9.4 %). Por último y confirmando esta tendencia, se observa que la estrategia **OD** logró identificar arquitecturas para : *Bosc*, *Brfl*, *Cisa*, *Lach*, *Oidi*, *Pema* con 25.1 %, 33.5 %, 32.6 %, 36.7 %, 26 %, 49.5 % respectivamente.

Por otra parte, la estrategia **OB**, solo logró identificar estructuras en *Brfl*, *Oidi* con 2.5 %, 6.7 % respectivamente, pero siendo inocuo al evaluar el resto de especies, de igual forma ocurrió con la estrategia **O** que identifico estructuras en *Brfl* con un 2.4 % pero es ausente en el resto de especies.

- *Mus musculus*: *Mumu*

Al evaluar las estructuras canónicas de *Mumu* en la gráfica **1-16** las estrategias **ODB**,

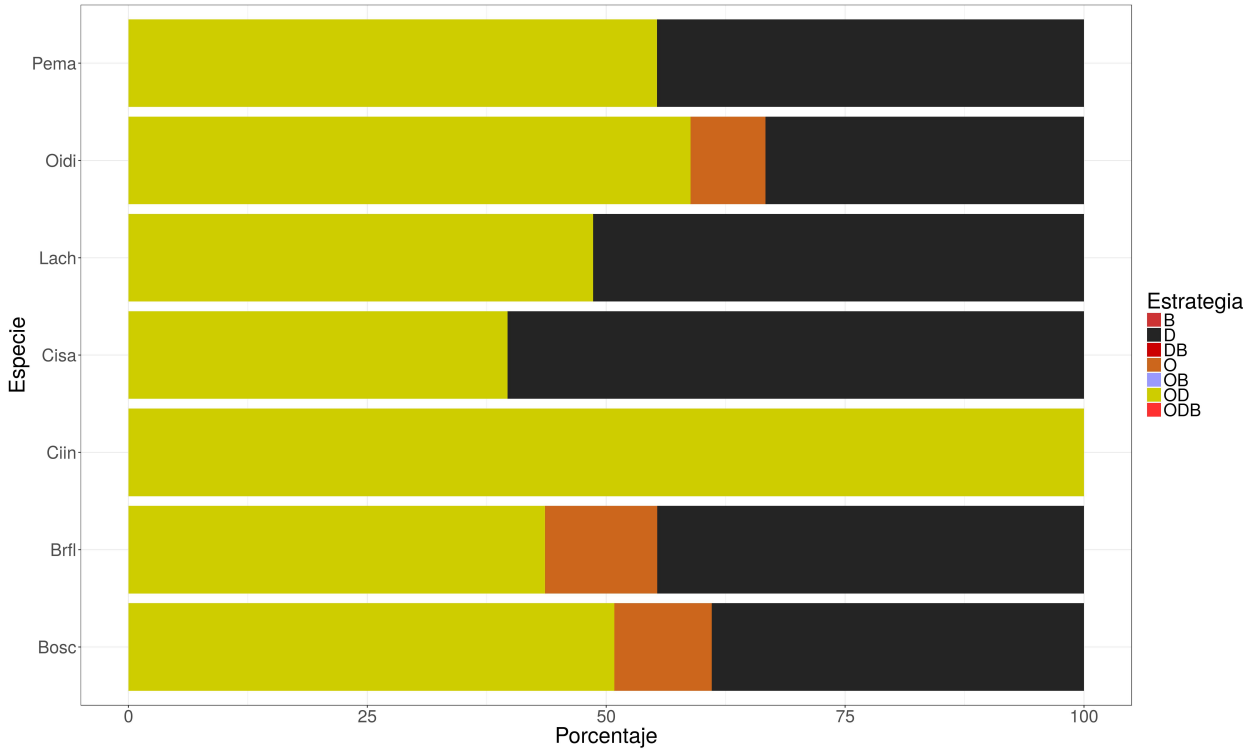

Figura 1-16: Porcentaje de las diferentes estrategias de anotación:**ODB** *Mus musculus* con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

DB y **OB** No lograron anotar ninguna proteína en las especies de estudio, mientras que la estrategia **O** no logro ningun tipo de anotacion en las especes *Ciin*, *Cisa*, *Lach*, *Pema* aunque en contraposición logro anotaciones en *Bosc*, *Brfl*, *Oidi* con 10.1 %, 11.7 % y 7.8 % respectivamente

El dato más revelador fue la estrategia **OD**, ya que fue significativa para la especie *Ciin* con un 100 %, al igual que otras especies pero con menores porcentajes: *Brfl*, *Cisa*, *Lach* con 43.5 %, 39.6 %, 48.6 % respectivamente, por último, se evaluo la strategia **D** y se encontraron anotaciones en *Bosc*(38.9 %), *Brfl*(44.6 %), *Oidi*(33.3 %) y *Pema* con 44.6 %

- *Homo sapiens*: *Hsa* Al momento de comparar el humano con los tunicados, se hace evidente, como lo muestra la grafica **1-17**, las estrategias de blast se vuelven menos efectivas como fue el caso de las estrategias **ODB**, **DB** ni **OB**, no lograron encontrar arquitecturas en ninguna especie. La mayoría de anotaciones con respecto a humano se dieron gracias a la combinatoria o uso de forma individual de las estrategias **D** y **O**, como es el caso de **OD**, la cual logró identificar arquitecturas, aunque con porcentajes bajos en las especies: *Bosc* con 35.3 % y *Ciin* con 28.2 %, Aunque fue evidente que el mayor éxito lo consiguieron las estrategias individuales de **D** , la cual logró anotar en las especies: *Brfl* (12.1 %), *Cisa* (33.3 %), *Lach* (31.2 %), *Oidi* (10.2 %) y *Pema* con

41.6 % .

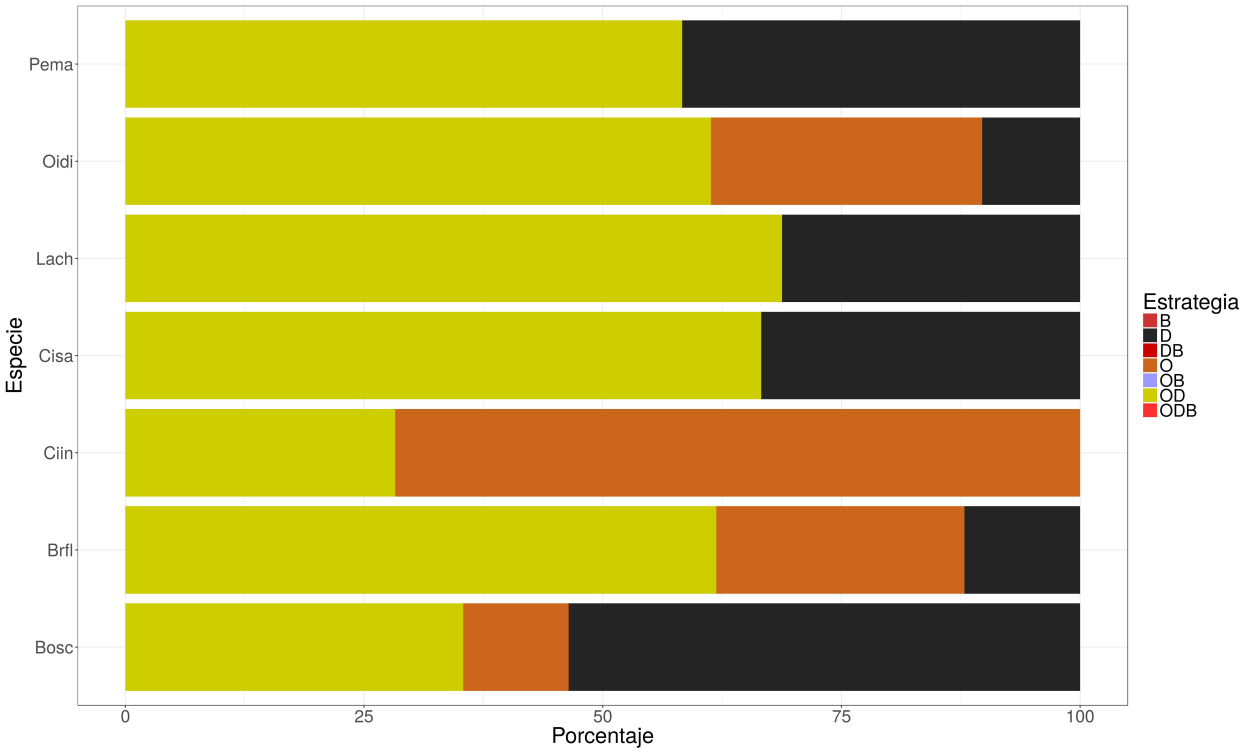

Figura 1-17: Porcentaje de las diferentes estrategias de anotación:**ODB** *Homo sapiens sapiens* con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

Las anotaciones se vieron complementadas por la estrategia logró identificar arquitecturas, aunque en bajo número, produjo anotaciones en las especies: *Bosc*, *Brfl* y *Oidi* con 11.0 %, 25.9 % y 28.3 % respectivamente, aunque vale aclarar que en las especies *Cisa*, *Lach* y *Pema* esta estrategia no tuvo efecto.

1.4.4. Obtención de genes asociados al sistema inmune

1.4.5. Pipeline de anotación automatizada

Para la totalidad de los genomas : *Lach* y *Brfl* *Bosc* (*Ciin* y *Cisa*) *Oidi* poseían predicciones de genes pero pocos genes anotados como se puede ver en la (Tabla**1-3**) en donde en la columna se registran los nombres de las especies, la columna dos el nombre de la base de datos de donde fue extraída la información, en la tres número de predicciones de genes según la base de datos, cuatro cuantos de esos genes cuentan con una anotación, cinco genes cuantos genes quedan asociados al sistema inmune. Es de rescatar que *Lach* y *Bosc* son las dos especies con mayor número de genes asociados aunque sea a un dominio del sistema

inmune y en contra de todo pronóstico *Oidi* a pesar de tener un genoma muy reducido duplica las predicciones de genes de genoma son reducidos como (*Ciin* y *Cisa*),

Tabla 1-3: Estado de anotación de los genes en las diferentes especies antes y después de la implementación de la pipeline

Especie	Base de Datos	Genes Base de datos	Anotados	Genes SII
Bosc	BSGP	30910	14731	763
Brfl	JGI	50817	NA	2353
Ciin	Ensembl	17153	795	942
Cisa	Ensembl	12172	298	837
Pema	Ensembl	13114	34386	893
Lach	Ensembl	45256	34386	4472
Oidi	OGB	18020	NA	432

[1] NA = No Disponible en las anotaciones proporcionadas por las paginas donde se descargaron los genomas

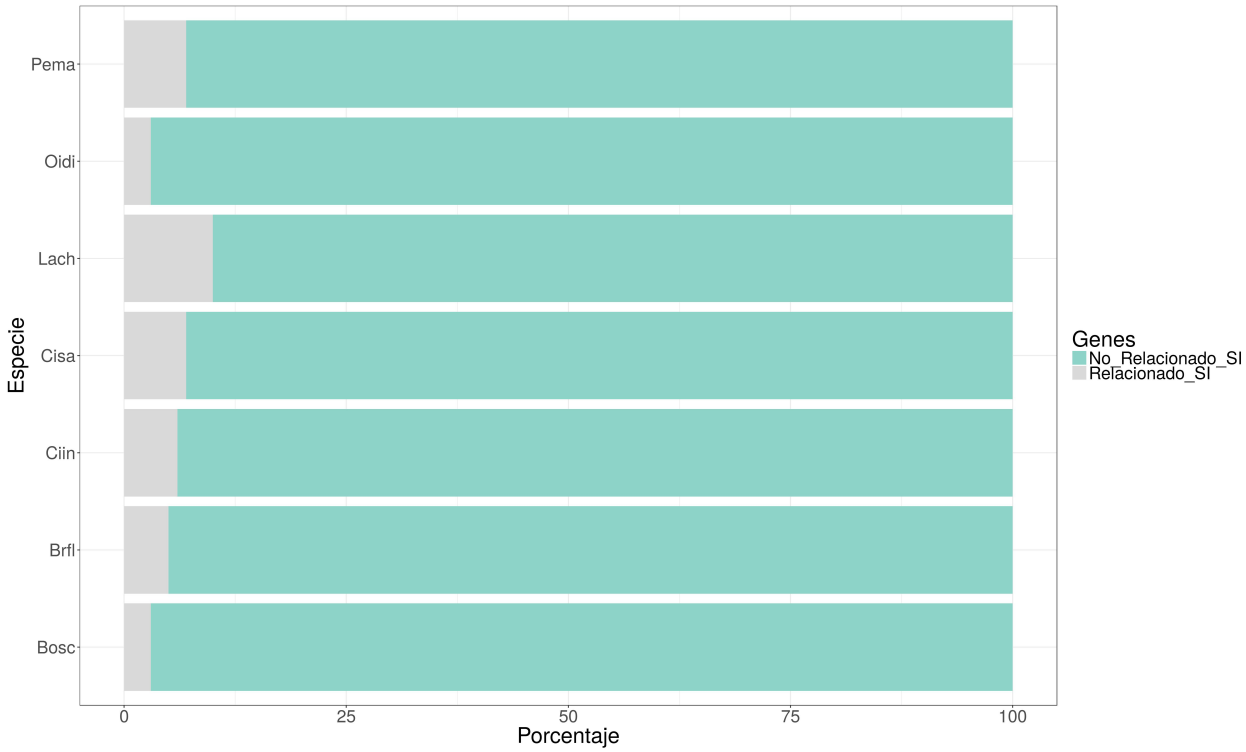

Figura 1-18: Proporción de posibles genes del sistema inmune con respecto a la totalidad de los genes predichos en cada una de las especies.

En la gráfica **1-18** se representa el porcentaje de genes del sistema inmune en la totalidad

de las anotaciones génica reportadas en cada una las especies, para esto se tomaron las predicciones génicas reportadas para cada uno de los genomas evaluados, tomando esto como el 100 %. Posteriormente se calculó cuál es el porcentaje representan los genes anotados mediante las estrategias ODB asociados al sistema inmune se puede observar que en *Lach* el porcentaje de los dominios del sistema inmune representa aproximadamente 10 % de los dominios, siendo el genoma con mayor anotaciones, mientras que el resto de especies muestran datos entre el 2 % y el 6 %, dos casos merecen ser rescatados, el porcentaje de 2,4 % de *Bosc* y la similaridad de porcentajes entre *Pema*, *Ciin* y *Cisa* con un 6 %.

1.4.6. Eficiencia del modelo

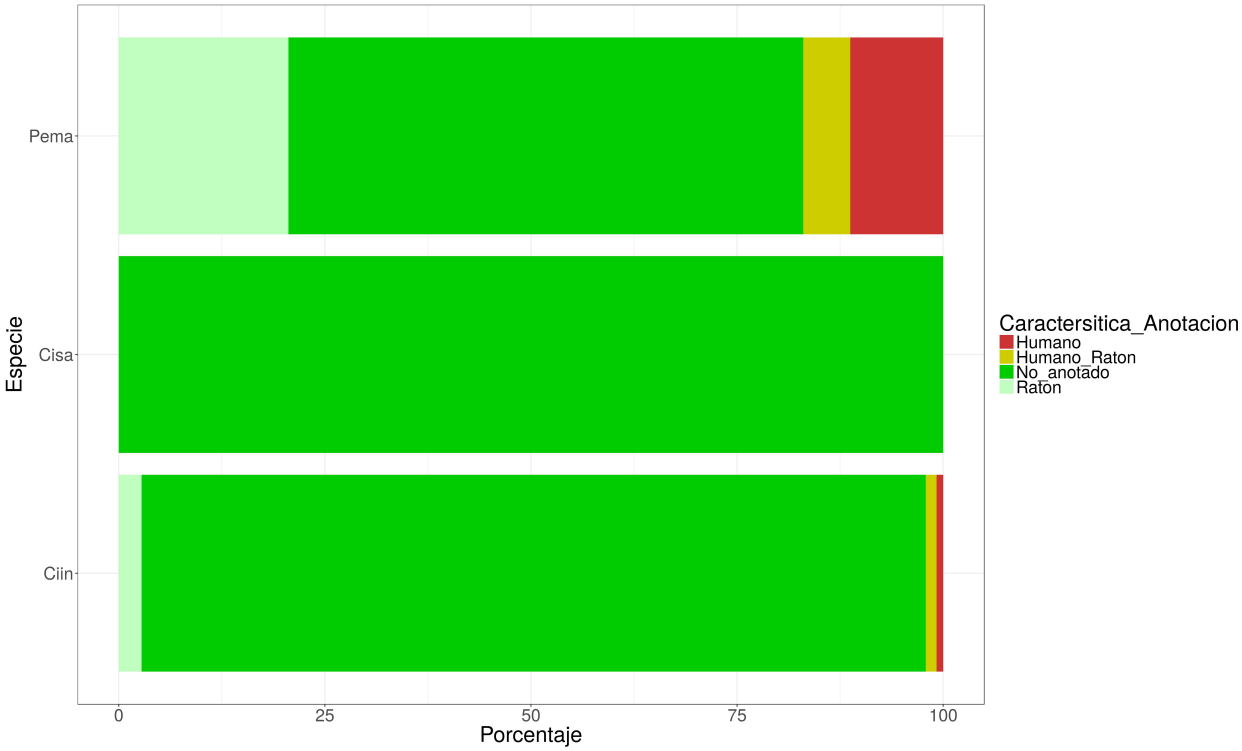

Figura 1-19: Esta gráfica muestra que porcentaje de las proteínas identificadas por la estrategia ODB que se encuentran con algún tipo de anotación en Ensembl, se evaluaron la proteínas encontradas en *Ciona intestinalis (Ciin)*, *Ciona savignyi (Cisa)* y *Petromyzon marinus (Pema)*que tuvieran arquitecturas similares y a su vez tuvieran asignado un nombre, dicha similaridad se evaluó tanto en en humano, ratón o ambas especies. Aquellas proteínas ausentes de nombre se catalogaron como No anotado.

Con el fin de establecer la eficiencia de esta metodología, se evaluó si existía concordancia entre el numero de nuestras anotaciones de genes putativos y las anotaciones de Ensembl.

Para esto, se tomaron los números de acceso de las proteínas de las especies **Cisa**, **Ciin**, **Pema** y **Lach** reportada por la estrategia ODB y posteriormente por medio de Biomart se extrajeron de Ensembl los nombres asociados a estas proteínas, y se clasificaron entre aquellas que no poseían nombres como "No anotadas" y aquellas que tenían nombre como "Anotadas". Debido a que las proteínas de referencia del sistema inmune provenientes de Innate data base pertenecen a proteínas de humano y ratón, se evaluó cuantos genes putativos tienen estructuras semejantes a ratón, cuantas a humano y cuantas estructuras fueron identificadas en ambas especies como lo muestra la figura **1-19**.

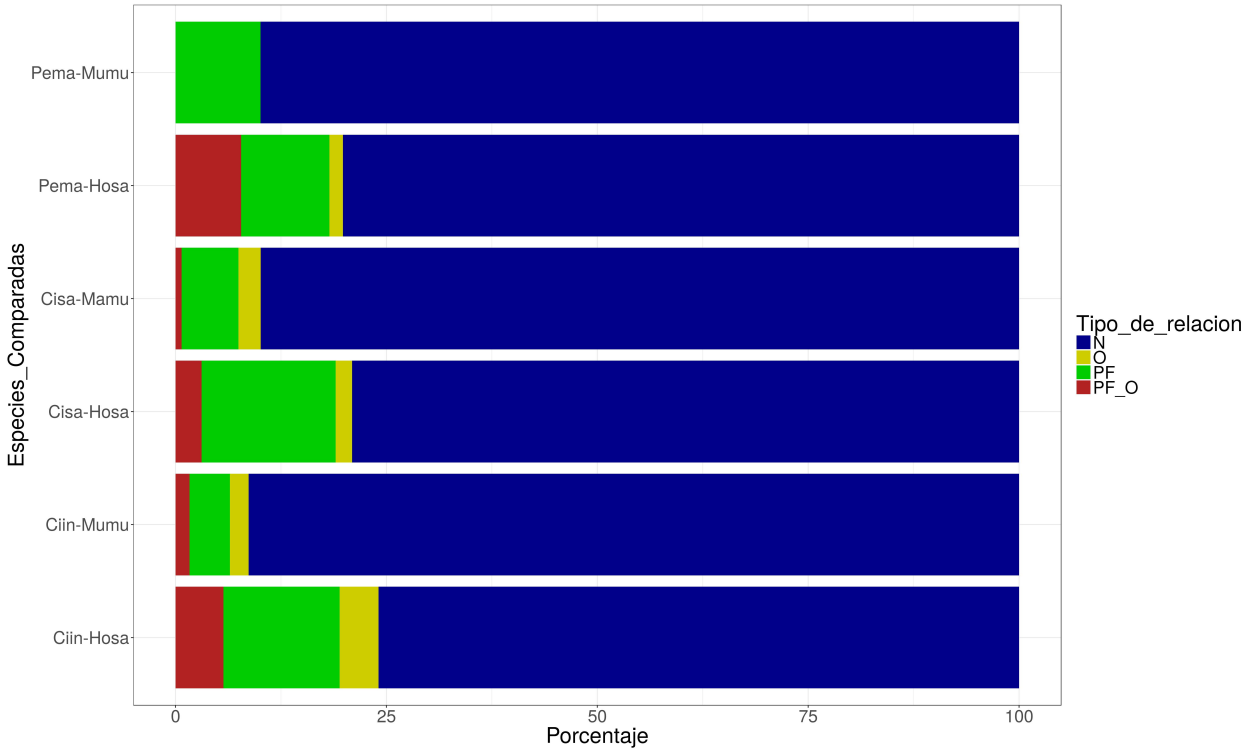

Figura 1-20: Gráfica muestra la relación que existe entre las proteínas predichas en **Cisa**, **Ciin** y **Pema** con las arquitecturas de humano y Ratón establecidas por la estrategia ODB contemplando relaciones de sub familias proteicas de Ensembl. La relación O_PF, denota relación de Ortología y de Familia Proteica, PF Familia Proteica, O Ortología y N No hay relación entre las dos arquitecturas.

Se observa en la figura **1-19** que **Ciin** de un total de 760 genes putativos identificados por la estrategia ODB, el 95 % no se encuentra anotado en Ensembl, el 1.3 % (10 proteínas) el nombre asignado en Ensmbl concuerda simultaneamente con los datos de ratón y humano, un 0.7 % (6 de proteínas) solo tienen arquitecturas canónicas similares a humano y un 2.7 % tienen similaridad exclusivamente con ratón. Por otra parte, **Pema**, de un total de 719 proteínas identificadas por la estrategia ODB, se encontró que 62.4 % no se encuentra anotado, un 5.7 % que corresponde a 41 proteínas anotadas en humano y ratón simultanamente, un

11.2 % corresponde unicamente a humano mientras que un 20.5 % que corresponde a 148 son proteínas anotadas en ratón. Por último, en **Cisa** de un total de 616 proteínas identificadas por la estrategia ODB, se encontró que 100 % no se encuentra anotado.
La figura **1-20** se observa que más 90 % de las proteínas en ratón provenientes de Ensembl no cuentan con ningún tipo de homología establecida con los tunicados, a pesar de esto, se observa un aumento tanto en el número de proteínas anotadas en tunicados cuya anotación por ODB han sido validadas con las relaciones de ortologia establecidas por Ensembl.
En comparación con los organismos modelo, los tunicados muestran conservadas las proporciones del porcentaje del genoma asociados al SI, de igual manera podemos ver que la estrategia ODB logra identificar de forma eficiente las familias proteicas propuestas por Ensembl pero es menos eficiente al replicar las relaciones simultaneas de ortologia y familias proteicas, y no es optimo al identificar unicamente la ortologia
En la gráfica **1-21**se muestra las relaciones de homología por familia proteica y ortología establecidas en las proteínas de los cordados basales y mamíferos sin tener en cuenta las subfamilias proteicas, en comparación con la gráfica **1-20**se observa un aumento significativo del número de proteínas correctamente anotadas llegando en la mayoría de los casos a ser mayor del 50 % reduciéndose de forma significativa el número de relaciones únicamente de ortología y un aumento en las relaciones simultáneas de Ortología y Familia proteica. en el único caso donde el número de proteínas sin una relación de homología aparente se dio en la evaluación de proteínas de *Cisa* con ratón.

1.4.7. Asociación de dominios en modelos asociados al SII en Tunicados

Debido a que gran parte de las ramas comunes de la evolución se ha demostrado que están constituidas a partir una serie de dominios ancestrales, Se decidió evaluar la frecuencia de cada uno de esos dominios: BIR, caderinas , CARD, CD36, Colágeno, Dominios lectina tipo C, EFG, IG, LRR, NACHT, Pkinasa, SCRCR, Sushii y TIR. Para objeto de medir las posibles variaciones de cada uno de estos dominios en la evolución del Sistema inmune, se evaluaron en cada una de las arquitecturas canónicas obtenidas mediantes las estrategias **ODB** (Figura **1-22**).
Los resultados muestran que la especie *Cisa* es la especie con mayor frecuencia de dominio asociados a arquitecturas del SII, seguido por *Lach* y *Brfl*, mientras que la especie que cuenta con menor número de dominios asociados a estructuras canónicas del SII es *Oidi*. Se observa que el dominio más predominante en la especie *Cisa* es EFG. En el caso de *Oidi* y *Bosc* se observa que el dominio que está altamente representado el dominio Pkinasa, Al evaluar la distribución de los dominios en *Brfl* hay un aumento en el número de dominios LRR en comparación con el resto de especies.
En la especie *Pema* se puede observar un aumento en los dominios de Colágeno, en el caso del dominios de colágeno también se observa este también observado en *Lach* y *Cisa*. Apesar

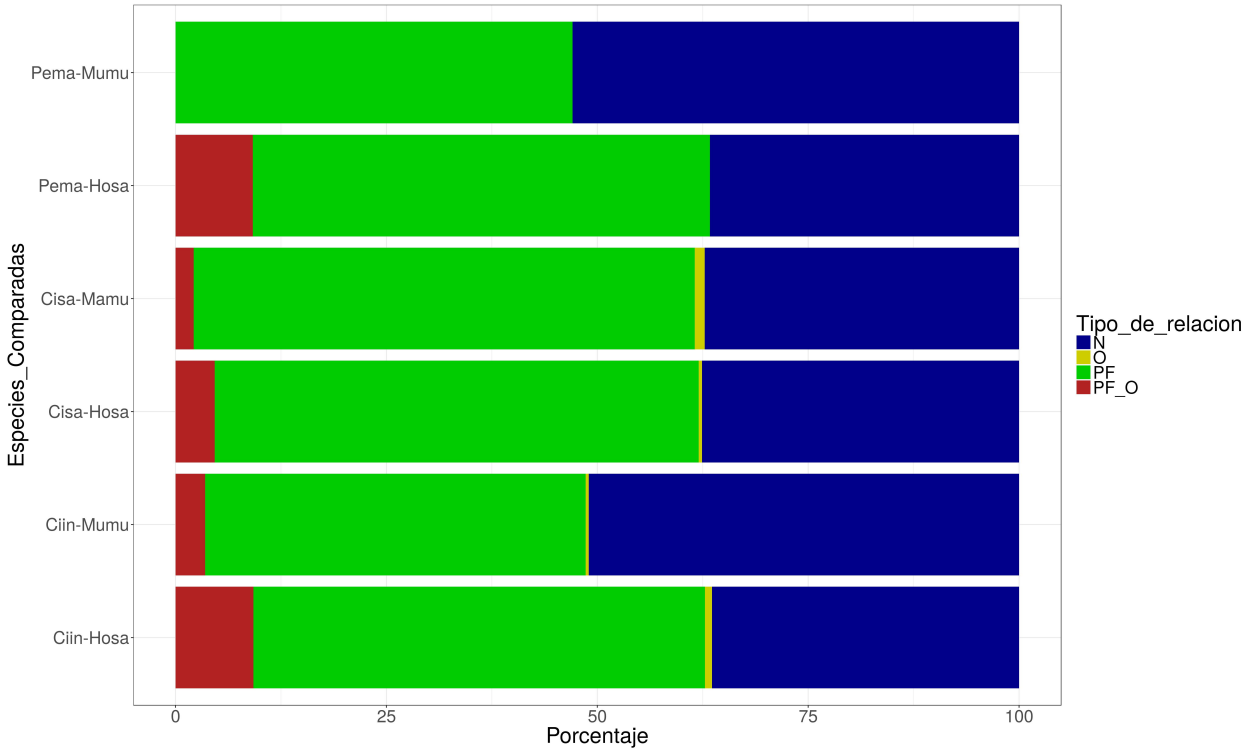

Figura 1-21: Gráfica muestra la relación que existe entre las proteínas predichas en **Cisa**, **Ciin** y **Pema** con las arquitecturas de humano y ratón establecidas por la estrategia ODB sincontemplar las relaciones de sub familias proteicas de Ensembl. La relación O_PF, denota relación de Ortología y de Familia Proteica, PF Familia Proteica, O Ortología y N No hay relación entre las dos arquitecturas.

que la duplicación de las inmunoglobulinas dio origen al SIA, se observa un aumento en el número de IG en las especies *Lach* y *Pema*.

Con el fin de tener un panorama general de la relación entre las dinámicas propias de cada dominio fundamental con referencia a sus dinámicas evolutivas, se evaluó qué porcentaje representa cada dominio en la totalidad de los dominios altamente conservados. Se encontró que la proporción de muchos dominios es constante a lo largo de la evolución como es el caso de las IG, que sólo advierten un cambio en los vertebrados mandibulados *Lach* y *Pema*. De igual forma la proporción de los dominios LRR es constante con un pequeño aumento en las especies *Pema* y *Brfl*, mientras que redujo su presencia en la especie *Oidi*. En otros casos como los dominios Pkinasa se ve una sobre representación de estos dominios en comparación con el resto en la especie *Oidi* y *Bosc*, por su parte el dominios de Colágeno y CARD tiene una mayor presencia en las Cionas, *Lach* y *Pema*, pero una pobre representación en *Bosc*, *Brfl* y *Oidi*. (Figura **1-23**).

El dominio BIR y CARD son dos dominios con historias particulares, ya que en el caso de

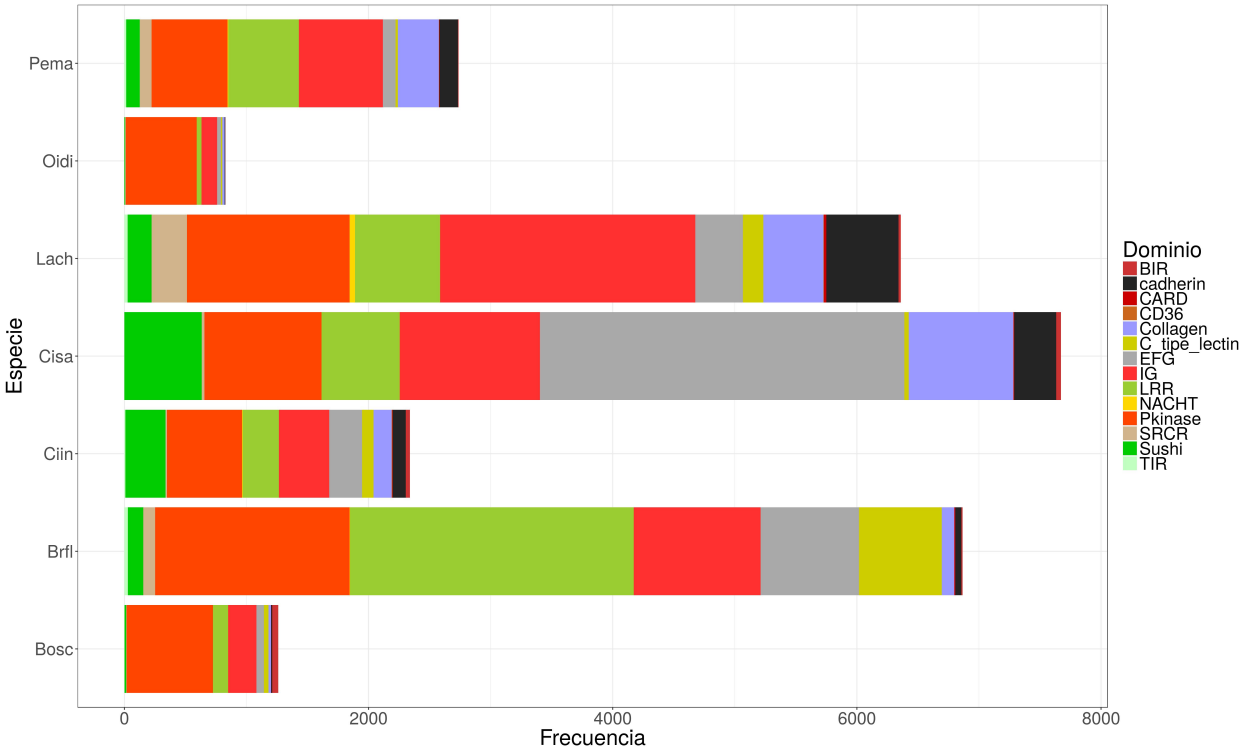

Figura 1-22: Frecuencia de dominios conservados son más preponderantes en cada una de las especies

BIR se ve una mayor representación en *Bosc* en comparación con el resto de especies,por su parte el dominio SUSHI tiene una notoria representación en la especie *Ciin*.

Moléculas de señalización

Para poder establecer la composición del SII, se decidió evaluar los diferentes módulos que componen la inmunidad, el Módulo señalización, Efector y de Reconocimiento. Este primero se destacan los dominios involucrados en vías de señalización intra y extracelular modulando las moléculas efectoras. En la Figura **1-24A** se muestran los dominios asociados a este módulo, se observa que el número de copias totales es bastante alto en *Cisa* y como ya se mencionó previemente las moléculas EFG son las más representadas, pero no solo son *Cisa* sino en todas las especies, se denotan los EFG como los dominios más importantes del módulo de señalización, hay dominios compartidos entre las Cionas y los vertebrados mandibulados que asu vez se encuentran subrepresentados en *Oidi*, *Brfl* y *Bosc*, entre estos dominios estan 7tm_1 y RCC1, por otro lado observamos que los dominios SUSHI está muy bien representado en *Cisa*, *Ciin* y *Lach*.
Al observar que proporción tienen estos dominios en el modulo modulo observamos **1-24B** el dominio LRR predominante en las especies *Brfl* y *Dare* con (37.6 %) y (25 %) respectivamente. mientras que en *Cisa* el sistema inmune se ve ampliamente representado por el dominio

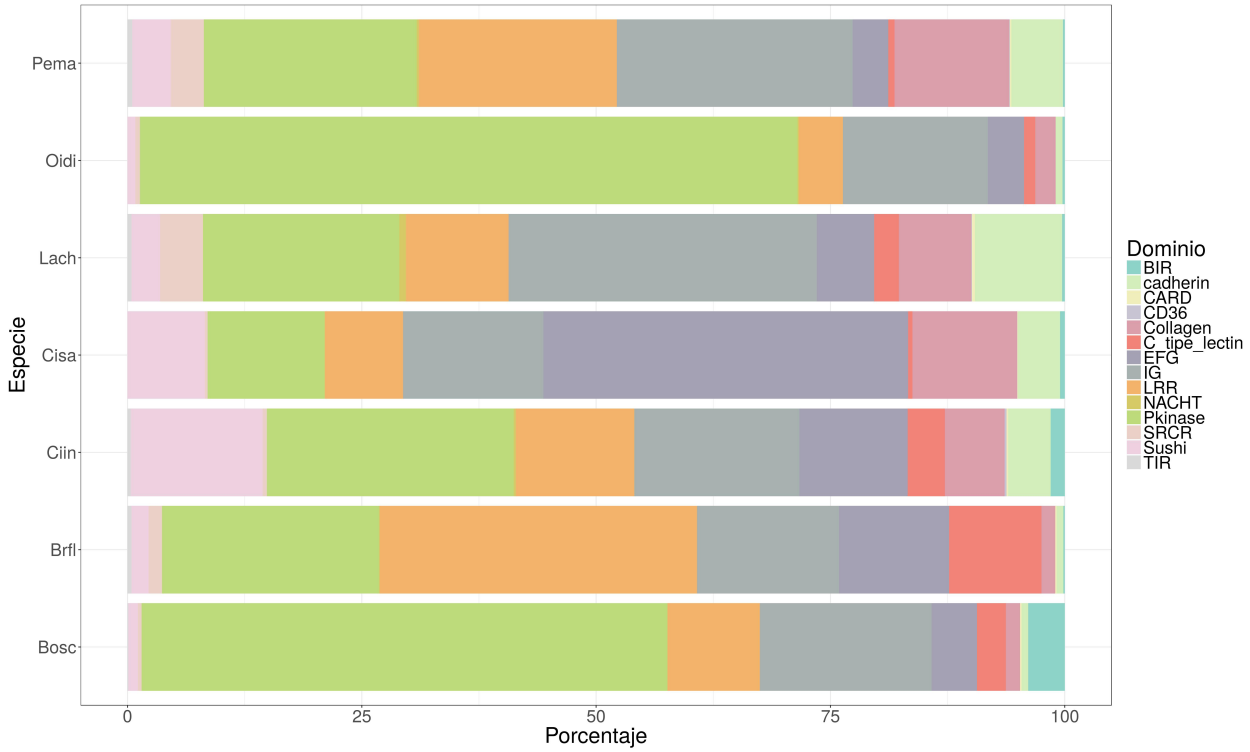

Figura 1-23: Proporción de Dominios asociados a genes anotados al sistema inmune por medio de la estrategia **ODB** en cada una de las especies de estudio.

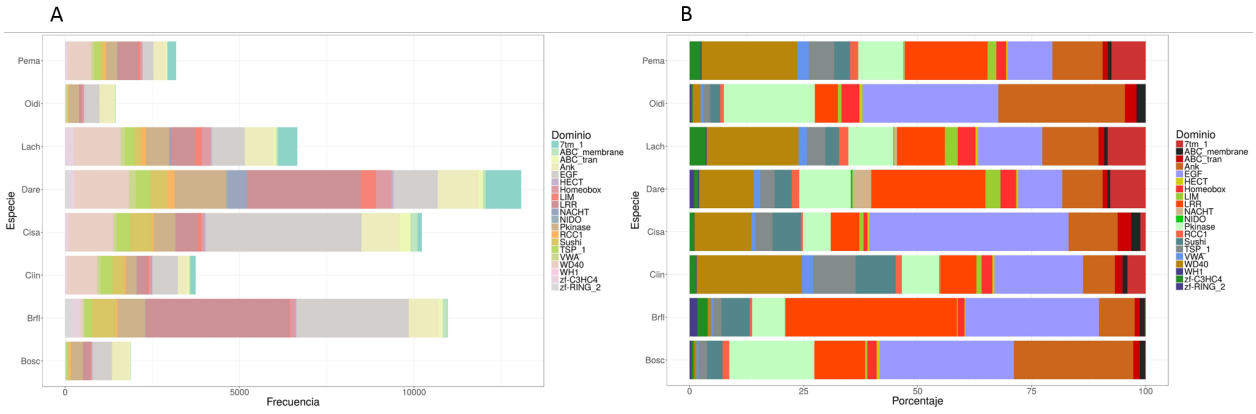

Figura 1-24: (**A**) Distribución de los dominios del módulos de señalización asociados a genes anotados al sistema inmune por medio de la estrategia **ODB** en cada una de las especies de estudio. (**B**) Distribución de las proporciones de los dominios asociados al Módulo de Señalización .

EGF con un (43.6 %), de igual forma se encontró que en *Lach* se vio fuertemente favorecido la aparición de moléculas de señalización donde se ve ampliamente representado el dominio WD40 on un (20 %). A pesar de verse favorecido un solo dominio, en comparación, se ob-

serva que en la especie *Pema* no existe un dominio predominante sino que por el contrario dos dominios WD40 (20.9 %) y LRR (18 %) son los mejor representados. por otro lado, se observa que la especie *Lach* los dominios EGF (14 %),Ank (12 %) y LRR (10 %) no tienen un papel principal en el sistema inmune de esta especie, pero representan dominios que juegan un papel importante. Sin embargo, a pesar de la existencia de dominios predominantes en el módulo efector, se observa que hay un set de dominios que siguen en importancia, y que es de este set de dominios donde aparecen estos dominios predomiantes, como es el caso de EGF donde no juega un papel protagónico como en *Cisa* en las especies *Brfl*, *Oidi* y *Bosc* y *Ciin* pero sí relevante, de igual forma el dominio Ank tiene un papel secuendario pero no relegado en *Pema*, *Oidi* y *Bosc*, mientras que WD40 asemeja el mismo caso en *Ciin* y *Pema*.

Moléculas de efectoras

Para distinguir entre las posibles variantes en las moléculas que permiten desencadenar una respuesta inmune posterior a un estímulo recibido [95] [84]. decidimos evaluar los diferentes dominios asociados a estas moléculas como se observa en la Figura **1-25A**, es esta gráfica se puede observar que hay una distribución homogénea de los dominios referentes a este módulo .

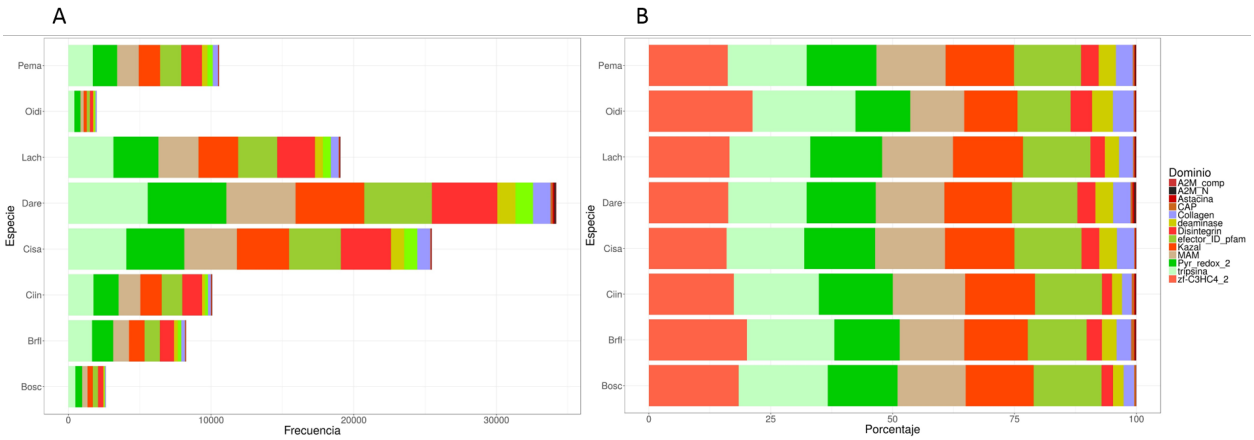

Figura 1-25: (**A**) Frecuencia absoluta de los dominios del módulos efector asociados a genes anotados al sistema inmune por medio de la estrategia **ODB** en cada una de las especies de estudio.(**B**) Distribución de las proporciones de los dominios asociados al Módulo efector.

El la figura **1-25***B* ilustra cómo está dispuesta las distribuciones de los dominios de este modulo, alli se puede observar que tanto para *Brfl* y *Oidi* los dominios zfC3HC4_2 , Tripsina son los mejor representados, mientras que en las especies *Bosc*, *Brfl*, *Ciin*, *Cisa*, *Dare*, *Lach* y *Pema* los dominios Desaminasa, dominio efector ID, Kazal, MAM, Pyr_redox, tripsina y Zf_C3HC4 son los dominios más representativos, denotando que no existe un dominio ampliamente favorecido.

Moléculas de reconocimiento

De igual se analizaron las frecuencias de los dominios asociados al módulo de reconocimiento, se puede observar en la figura **1-26A**, una alta frecuencia del dominio fn en *Cisa*, *Pema* y *Lach*, de igual forma vemos que el dominio WD40 se encuentra representado en las Cionas y los vertebrados mandibulados pero ausente en el *Brfl* y el resto de tunicados. Cabe también señalar que en el módulo de reconocimiento es la primera vez en donde *Lach* y *Cisa* tienen frecuencias similares.
En la Figura **1-26B** se observa que el resultado más impactante que surge de los datos es que en la especie *Bosc*, *Oidi* se observa que la evolución del sistema inmune en esta especie favorece las proteínas con el dominio Pkinase (40 %) y (49.7 %), mientras que en especies como *Ciin*,*Cisa*, *Dare*, *Lach* y *Pema* se encuentra entre los más relevantes aunque no juega un papel principal. De igual forma se destaca el LRR en *Brfl* y *Dare* con un (46.4 %) y (22 %) respectivamente y en *Ciin* se observa que la evolución del sistema inmune en esta especie favorece las proteínas con el dominio WD40 (26.5 %).

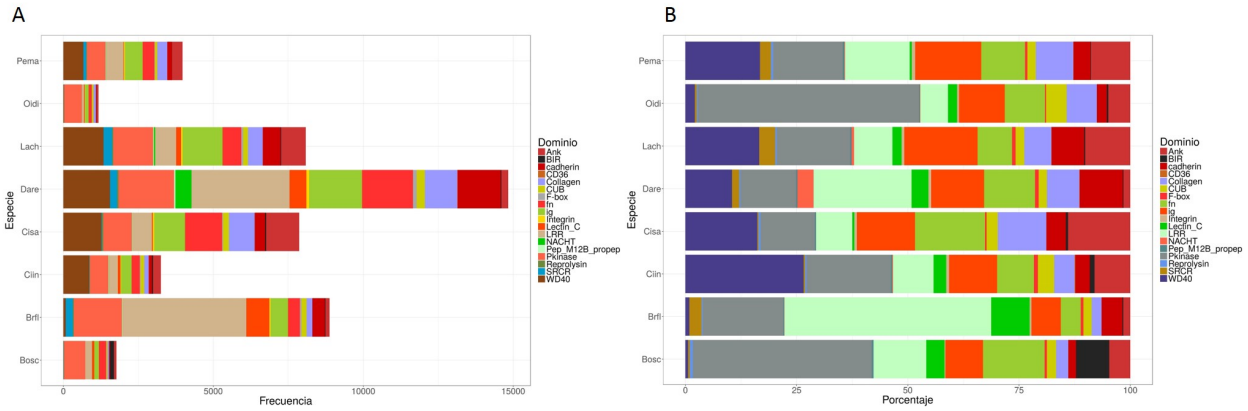

Figura 1-26: (**A**) Frecuencia de los dominios del módulo de reconocimiento asociados a genes anotados al sistema inmune por medio de la estrategia **ODB** en cada una de las especies de estudio.(**B**) Distribución de las proporciones de los dominios asociados al Módulo de reconocimiento.

Al contrario, se puede ver en la figura que en *Lach* que existen dominios que en conjunto tiene un papel protagónico como los son WD40 (16.5 %),ig (16.5 %),Pkinase (16.4 %), seguido por otro set de dominios Ank (10 %), LRR (8.6 %), fn (7.7 %), cadherin (7.3 %) juegan un papel complementario entre sí, caso similar a *Cisa* donde los protagónicos son el conjunto compuesto por WD40 (16 %),fn (15.7 %),Ank (13.9 %),ig (13 %),Pkinase (12 %) y los secundarios están representados por Collagen (10.9 %) y LRR (8 %). El último caso en el que esta diversificación de dominio se da es en la especie *Pema* donde los dominios más representativos están representados por WD40 (16.7 %),Pkinase (15.6 %),ig (14.8 %),LRR (14.5 %) mientras que en los secundarios están fn (9.8 %),Ank (8.8 %) y Collagen (8.4 %).

1.5. Discusión

En este trabajo observamos varios problemas que se encuentran cuando se desea trabajar con el sistema inmune. Uno de los grandes problemas fue tratar de definir un sistema que por su naturaleza en si es difícil de definir ya que sólo basados en su concepto más amplio que indica que debe mantener la integridad genética y fisiológica del organismo entonces se hace más difícil definir que tipo de componentes hacen parte del sistema inmune y cuales no. La manera de apoyo para conocer esos componentes es altamente dependiente de las bases de datos existentes pero debido a la importancia médica que tiene este sistema, los registros están altamente dependientes de estudios sesgados al estudio de especies modelo como ratón o concentrados en recopilar información del humano. Es por ello que en este trabajo nos centramos en dos fuentes principales de información del sistema inmune como fueron las bases de datos `InnateDB` [18] e `Insect Innate Immunity Database` (IIID) [20] que fueron estratégicamente seleccionadas ya que en sus registros de genes y de proteínas la función asociada a la inmunidad innata se encuentra experimentalmente validada permitiendo así partir de una fuente de datos donde la función inmune está claramente definida. Con estas bases de datos fue posible identificar un grupo de genes putativos asociados al SII de las especies de estudio los tunicados. Para estas especies existe poca información sobre el SI y una versión cercana a la inmunidad de estos organismos se puede resumir en el trabajo [122].

Al observar la **1-10** se observan las principales fuentes de anotación utilizadas en esta aproximación. Note que por otro lado al comparar las tablas de presencia y ausencia de dominios en las arquitecturas canónicas (**1-1**) y los dominios que se rescatan luego de las estrategias de anotación (**1-2**) se observa que este modelo es bastante astringente, ya que pro ejemplo bases de datos como SUPERFAMILY pasa de tener 897 a solo 448 después de aplicar las estrategias de anotación, casi la mitad de los dominios, lo cual se evidencia de nuevo al comparar los dos datos en la base de datos de Pfam, se pasa de haber tenido 3615 a 1195 después de haber aplicado las estrategias. aunque sin duda alguna los caso más dramáticos están los ya mencionados con el resto de bases de datos y el caso de TIGRFAMs que pasa de tener 344 a 36, aunque este tipo de resultados puede ser inquietante, debemos señalar que los retos que genera la anotación de estructuras proteicas filogenéticamente distantes hace imposible flexibilizar los filtros de anotación.

Al observar la **1-10** se destaca la ausencia de dominios de las bases de datos Phanter, la cual fue la base con el mayor número de tractos de CDS que codificaran para dominios asociados a estructuras canónicas del sistema inmune , al igual que bases de datos como Smart, Prosite y PIRSF, esto se puede deber a que las arquitecturas canónicas pueden estar representadas por un solo dominio lo que hace que sean filtradas y no tenidas en cuenta en las diferentes estrategias, Posteriormente a tener identificados los dominios por cada una de las bases de datos, afrontamos uno de los problemas apasionantes del sistema inmune, la alta combinatoria de los módulos fundamentales del sistema inmune se convirtió en el reto más demandante de esta tesis, debido a que la alta tasa de duplicación dentro del gen de

estos tractos de CDS que codificaran para dominios y su continuo reordenamiento llevaron a desarrollar tres diferentes metodologías que resultaron ser exitosas al combinarlas entre ellas.

Este reordenamiento acompañado de la reducción en la cantidad informacion tambien se ve reflejado en el número de genes hallados mediante las estrategias ODB, ya que sorprendentemente se evidencia la disminución de la representabilidad de los genes asociados al SII en la especies *Bosc* (como se evidencia en la gráfica **1-18**) un porcentaje similar a los de la especie *Oidi* en esta última especie se puede explicar este fenómeno debido a los diferentes procesos de reducción del genoma que ha sufrido esta especie, lo que sería una aproximación al número de genes que se evidenciaron, además de esto, se espera que aparezcan proteínas con un bajo número de dominios se puede dar por la compactación del su genoma evidenciado en la reducción de sus regiones intergénicas, [10], por el contrario, en una especie de tunicado colonial como *Bosc*, el cual no presenta una reducción del genoma y no reducciones génicas, se esperaba que los retos de evitar las quimeras por medio del alloreconocimiento le diera una importancia mayor a los genes de la inmunidad, [116], factiblemente este fenómeno está relacionado genes huérfanos o innovaciones evolutivas cuya diferenciación no es objeto de esta tesis, genes cuya función sea específica a ese grupo y no sea rastreada por medio de comparación de arquitecturas con otro grupos. fenómeno que explicaría que especies no coloniales como *Ciin* y *Cisa* tengan números similares en cuentro a identificación de proteínas por comparación, sumado a que las Cionas poseen una la alta densidad génica debido a lo compactado que se encuentran los introns, pero aun asi se ve una baja pérdida específica de genes presentando regiones conservadas fácilmente rastreables a pesar de la pérdida masiva de genes reportada en la literatura [10]. Por otra parte, *Brfl* presenta una serie de expansiones del genoma por medio de duplicaciones del genoma tanto locales como globales , observando los resultado obtenidos en *Lach*, se puede evidenciar el hecho de que los genomas de teleósteos han sufrido un evento de duplicación de todo el genoma, así como eventos de duplicación de genes individuales. En donde dominios como las IG se han visto fuertemente favorecidas. Todo esto sugiere que hay un conjunto aproximado de 700 genes fundamentales en el SII de tunicados de los cuales más de 400 son de alta importancia ya que se encuentran hasta en genomas altamente reducidos.

Al momento de evaluar las diferentes estrategias de anotación encontramos que las por otra parte se pudo observar que las estrategias de **O**rden y **D**esorden fueron las más efectivas para encontrar homólogos remotos tanto de humano a tunicados como de insectos a tunicados, lo cual confirma que hay un grupo de genes altamente conservados en la inmunidad, cuyas relaciones filogenéticas pueden ser trazables desde insectos hasta los mamíferos, el éxito de estas dos estrategias también indica que minimizar la variable de número de repeticiones y concentrarse en la arquitectura genética basada en arquitecturas gold standard es un método bastante novedoso que soluciona en parte las desventajas del uso de Blast en homólogos remotos. en el momento de comparar es resaltable que el grupo Hymenoptera, compuesto en nuestro estudio por *Nasonia vitripennis* y *Apis mellifera* y el grupo Diptera compuesto por

: *Anopheles gambiae* y *Drosophila melanogaster* , es clave señalar que el grupo intermedio entre los icetos evaluados y los cefalocordados, *Acyrthosiphon pisum* , un grupo a los insectos homopteros no presenta tantas homologías con blast como si los tienen los grupos ancestrales Hymenoptera, compuesto en nuestro estudio por *Nasonia vitripennis* y *Apis mellifera* y Diptera compuesto por: *Anopheles gambiae* y *Drosophila melanogaster* , sino que por el contrario tenga un comportamiento similar a los mamíferos donde la identificación con blast fue la estrategia menos efectiva. miesntras que en *Nasonia vitripennis*, *Apis mellifera* y *Anopheles gambiae* se evidencia al estrategia basada en blast , aunque se observa que la estrategia de **DB** tuvo cierto exito en el grupo Hymenoptera en comparación con los tunicados. aunque al momento de rastrear las arquitecturas canónicas fueron bastante eficiente las estructuras ordenadas de los mamíferos.

Al observar el comportamiento en los dominios más conservados, se observa un aumento significativo en la frecuencia del dominio EFG en la especie *Cisa*, debido a que el dominio EFG está involucrado en procesos de reconocimiento de ligando, adhesión en las proteínas tipo P-selectinas, procesos inflamatorios e interacciones proteinas proteínas, siendo de vital importancia en la unión a células en proceso de apoptosis citewouters2005evolution, al momento de evaluar este dominio en el contexto de los dominios asociados al módulo de señalización, se observa que es el dominio con mayor número de copias entre los dominios de señalización, superando a dominios clásicos como LRR, sugiriendo que en los organismos como los tunicados, donde el tejido epitelial juega un papel fundamental, ya que aparte de ser la primera barrera inmunológica ya que cumplen funciones secretoras y fagociticas lo que explicaría las múltiples copias de dicho dominio en *Cisa* [84].

Al evaluar los dominios efectores se denota una composición equitativa en la presencia de los diferentes módulos relativos a los procesos efectores del sistema inmune, se observa una tendencia altamente conservada en términos de proporción de los diferentes dominios de este módulo, lo cual es de esperarse ya que previo a la aparición del SIA gran parte de la respuesta inmune recae sobre estas moléculas, por lo que era de esperarse que fuera un sistema con fuerzas de presión positivas que mantuvieran la efectividad de este módulo, en contraposición se espera que haya un aumento en la frecuencia de los dominios sin cambiar las proporciones con el objetivo de que prevalezca la plasticidad de estas moléculas y por ende su diversificación, un resultado sorprendente es que se esperaba un cambio significativo en la proporción de los dominios tipos lectinas en el paso de vertebrados no mandibulados a los mandibulados, ya que en este segundo la aparición del SIA implicaría un "alivianamiento" en mantener una respuesta mediada por lectinas ya que sería compensada por los anticuerpos, [95] al ver nuestros resultados evidenciamos que este proceso no ocurre en *Lach*, pero si en *Pema*, lo que nos indica que la efectividad y ventajas que tiene las lectinas no se ve afectado por la presencia del SIA.

Por otro lado es evidente la importancia de los dominios Pkinasas en *Oidi* y *Bosc* indican que los procesos de reconocimiento citoplasmáticas se ven favorecidos, ya que se logra ver tanto en la gráfica **1-22** como en **1-26** un aumento en los dominios de la vía de señalización del

SII, Pkinasas, dicho dominio está involucrado en la activación de la expresión de genes que pueden terminar favoreciendo sus productos procesos inflamatorio mediado por moléculas efectoras o en la muerte celular con el proceso de apoptosis, pero de igual forma es intrigante proque en la especie *Brfl* juega un papel secundario sin mayor relevancia, siendo de esta forma la única excepción [79].

Al igual que en el Módulo de señalización se observa la predominancia de ciertos dominios como Pkinasas, LRR y WD40 en este módulo, que relegan a posiciones menos visibles al resto de dominios, sin embargo al igual que en el módulo de señalización, se destaca un conjunto de dominios que mantienen cierto protagonismo y que dependiendo de la especie asumen mayor o menor relevancia, como lo son los dominios de Colágeno, Ank, fn y Ig. Especies como *Lach*, *Cisa* y *Pema* muestran que es una diversidad de dominios y no un dominio protagónico los dominantes en este módulo

Por otro lado, debido a la diversificación de los dominios de reconocimiento se diversificó las rutas de señalización favoreciendo la aparición de módulos completamente diferentes donde en cada especie favorece módulos dominados por diversos dominios sin una tendencia aparente, lo cual se puede observar en dicho modulo, ya que a diferencia de los módulos de Reconocimiento y Efector encontramos casos como *Cisa* donde el dominio EFG representa casi la mitad de los dominios de este módulo, caso similar de *Dare* donde casi el 40 % de los dominios se encuentra representado por LRR. De este último dominios resalta ante todo que LRR es un dominio relevante en los cefalocordados y Vertebrados y mientras que en los tunicados juega un papel secundario, De igual forma se evidencia que EGF es un dominio relevante en las especies cefalocordado, tunicados y *Lach* pero tiene un papel menos importante en los primeros estadios de la aparición de los vertebrados como se observa en *Dare*,*Pema* juega un papel se ve opacado *Pema*.

Se observa un aumento en la frecuencia de dominios de *Cisa* en relación con las otras especies, siendo la especie con mayor número de dominios (Figura **1-22**), pero al evaluar el número de genes asociados al SII se observa que tiene un número menor de genes que *Ciin*, lo que indica que hay un aumento en el número de copias de dominios por gen en comparación con *Ciin* sin que esto implique el aumento en el número de genes asociados al sistema inmune, dicho aumento de complejidad es visto en *Brfl*, *Lach* y *Pema* donde cada proteína tiene una expansión de dominios, pero en contraste está acompañada por una expansión de la información genética que deriva en un aumento en el número de genes.

En el caso de *Brfl* donde se evidencio un aumento en el número de dominios LRR en comparación con el resto de especies, acompañado de un aumento en el número de genes creemos que este aumento de dominios señalizadores se puede deber no solo a una complejización de los genes ya existentes con duplicaciones de dominios, sino que en sí mismo a una expansión de los genes ya existentes como es el caso de los TLRs, ya que a su vez se encuentra una expansión de los dominios tipo C lectina los cuales también están presentes en estos receptores.

Sobre los dominios IG se observa un descubrimiento interesante, aunque la expansión de la

inmunoglobulinas en vertebrados mandibulados dio paso a la expansión del SIA, se observa que en *Lach* hay un leve aumento de la proporción de dominios IG lo que puede indicar que las dos rondas de duplicación no solo permitieron la aparición del SIA sino que pudieron estar involucradas en duplicación de genes asociados al SII o a los dominios que los componen. Por otro lado vale resaltar que los dominios de Cadherina y Colágeno empiezan a tener mayor relevancia en las Cionas y en los Vertebrados mandibulados, pero sub representados en *Bosc*, *Brfl* y *Oidi*, lo que puede indicar eventos de expansión de ese dominio recientemente en dos puntos del árbol evolutivo, en un ancestro de los vertebrados mandibulados posterior a la división a los Cefalocordados y a la aparición de los tunicados, y un evento específico en la rama de las Cionas

Se identificó que para las especies más basales de cordados de ensembl el nivel de anotación genómica es casi nulo, como se demuestra en la figura **1-19** donde tanto **Ciin** como **Cisa** más del 90 % de las proteínas carecían de nombre, por otra parte se identificó que varias de las proteínas asociadas a la sistema inmune del ratón fueron identificadas con nombre, caso contrario con humano que tanto en **Pema** como en **Ciin**mostraron números menores que Raton. de igual forma se hizo evidente que proteínas anotadas compartidas tanto por humano como ratón en el sistema inmune son reducidas. Es por esto que el papel de esta nueva metodología cumple un papel fundamental y es plantear una forma efectiva que a partir de las arquitecturas canónicas se puedan establecer relaciones de ortología en especies tan lejanas.

por otro lado en la figura **1-20** se evidencia que las relaciones de ortología no sobrepasan el 15 % de las proteínas identificadas y que contrario a lo esperado, las familias proteicas no se ven reflejadas en las relaciones de ortología, lo cual se puede explicar debido a que conformar estas relaciones se dio por medio de dos metodologías diferentes, una basada en blast recíproco como lo son las relaciones de ortología y el establecimiento de las familias proteicas que está basado en HMM, lo cual puede explicar como este último fue más efectivo al momento de establecer relaciones de homologías lejanas como los son mamíferos tunicados ya que o HMM están construidos con base a todas las proteínas del árbol de la vida. Por otro lado en la figura **1-21** identificamos que la metodología en la mayoría de los casos logra identificar más del 50 % de proteínas correctamente, esto ocurre cuando se evalúa la familia proteica más grande y no se comparan las subfamilias proteicas de ensembl. Factiblemente las dificultades que tiene esta metodología para la identificación de subfamilias proteicas se debe a la evaluación por listas de la metodología ODB, la cual no tiene en cuenta el número de repeticiones de cada dominio y esto puede ser fundamental a la hora de clasificarlas en estas subfamilias.En cuanto a las proteínas No anotadas, es válido aclarar que muchas de las proteínas en ensembl poseen anotaciones como AMBIGUOUS' o 'UNKNOWN' que evidentemente no va a poder ser clasificada como Anotada correctamente al momento de evaluar las estructuras.

, lo que lo hizo un reto interesante pero de una alta complejidad, consideramos que fue un acierto el uso de estructuras canónicas basada en para anotar el sistema inmune, basada en

encontrar tractos de CDS que codificaran para dominios asociados al sistema inmune. Después de haber podido definir las estructuras canónicas y haber delimitado las los dominios canónicos reales asociadas al SII, nos vimos enfrentados a la amplia diversidad de aproximaciones que poseen las diversas bases de datos asociadas a Uniprot , dichas estrategias divergen en la metodología para curar y generar HMM, lo que implicaba que las diferentes arquitecturas definidas como gold standard no podían ser comparadas si provienen de bases de datos diferente, evidenciando que el cruce de información entre estas bases de datos sigue siendo un problema sin resolver.

2 Capitulo 2: Pipeline de anotación de genes asociados al sistema inmune en tunicados Didemnum vexillum

2.1. Introducción

2.1.1. Arquitectura Génica

Para poder comprender los diferentes algoritmos de predicción de genes de novo es necesario entender las partes que componen un gen o lo que denominamos arquitectura genética. Este conjunto de estructuras incluyen tanto secuencias que codifican para dominios de proteínas así como regiones que no son traducidas, siendo estas últimas de alta importancia para la regulación de la expresión de un gen debido a la presencia secuencias reguladoras o incluso regiones asociados al procesamiento de los intrones. La ubicación de estas regiones no está restringidas a los extremos UTRs del gen sino que pueden estar ubicadas en intrones o en alternantes señales de poli-adenilación de los extremos 3. Las configuraciones de dichas regiones varía tanto en especies como en tipos celulares al existir en algunos casos sobrelapamiento o entrecruzamiento de regiones que permiten múltiples cambios en la estructura de los exones o de las UTRs y por tanto hacen que la arquitectura sea casi especie específica. [48]
A pesar de las combinatorias casi que únicas de cada gen, se pueden establecer la presencia de 9 regiones que pueden estar o no presentantes en la arquitectura de un gen , estas son: 1) región promotora , 2) Enhancers 3) La región 5 UTR 4) Exones 5) Intrones 6) Codón de Parada 7) región 3 UTR. [48]
A continuación explicaremos brevemente la configuración de cada una de ellas:

- `Promotores`: Los Promotores son una región del gen, ubicada normalmente corriente arriba o corriente abajo del sitio de inicio de la transcripción (normalmente representado por la secuencia ATG). En dicha región se unen factores de transcripción, que facilitan el reclutamiento tanto de la RNA polimerasa como de elementos próximos al promotor que ayudan a la unión de la misma, como por ejemplo, los elementos de reconocimiento B. Es por esto el nombre de promotores ya que estas secuencias promueven la transcripción. Se diferencian dos tipos de promotores, los conservados los cuales se

encuentran enriquecidos por secuencias repetitivas de Timina y Adenina o Caja TATA y los promotores variables los cuales tiene como característica estar enriquecidos por secuencias CpG. [48]

A lo largo de la evolución se ha observado que los promotores complejos pueden estar tan especializados en eucariotas, que pueden promover una variante genética requerida no sólo al tipo celular y además variar dependiendo de la etapa de desarrollo. Toda esta complejidad, en la conformación de los promotores, es una de las claves de la compleja gama de patrones de expresión necesarios para la diferenciación celular, dicha innovación ha sido clave en el desarrollo de organismos complejos en los cordados[57]. Pero a pesar de esta complejidad, se sabe que los promotores por sí mismos no pueden actuar solos, sino que requieren de una serie de regiones reguladoras adicionales conocidos como enhancers.

- `Enhancers`: los enhancers son secuencias que se encuentran corriente arriba o abajo de la región promotora,. La principal función de estas secuencias es controlar la eficiencia y tasa de transcripción de los genes al interactuar con los promotores. Dos características hacen a estas secuencias especiales 1) su función moduladora actúa de forma selectiva ya que solo pueden activar promotores enlazados en cis, y 2) puede encontrarse tanto cerca como distante a la región promotora e igual cumplir su función. [57]. Existen otro tipo de reguladores como los factores de transcripción los cuales se unen a regiones promotoras o potencializadoras para poder activar o inhibir la transcripción de un gen. Se ha observado que a la par de la aparición a lo largo de la evolución de promotores complejos han evolucionado enhancers entre los que se encuentran los aisladores, activadores o represores (silenciadores)[8].

- `Región 5 UTR`: la región 5 UTR es una región corriente arriba del sitio del sitio de inicio de la transcripción, la cual es transcrita más no traducida en los genes codificantes de proteínas, su función es controlar el inicio de la trascripción debido a que induce respuestas específicas a condiciones ambientales y señales celulares, el control de estas respuestas favorecen la expresión de variantes génicas requerida acordes a la etapa de desarrollo y el tipo de célula. A largo de la evolución se ha observado que este control ha sido una ventaja debido a que es uno de los requisitos básicos para la producción patrones de expresión complejos necesarios para el desarrollo de organismos complejos. Debido a la alta complejidad de esta región se han divido un compendio de estructuras entre las cuales se encuentran estructuras secundarias y la región d deformación del cap 5. Estas estructuras secundarias afectan directamente la tasa de transcripción de genes en donde UTR cortas, con baja cantidad de CG, ausencia de codones prematuros de parada y sin estructura definida son característicos de genes con alta tasa de transcripción, mientras que por el contrario los genes con baja producción tienen UTR más largas y son ricos CG. Otra de las características de esta región está sujeta a splicing alternativo y variaciones que permiten que una sola UTR genere una

amplia gama de expresión génica en solo un locus [91][57]. La estructura secundaria también se caracteriza por la presencia de motivos reguladores como los sitios internos de entrada al ribosoma o Internal ribosome entry sites (IRES), los cuales pueden iniciar traducción de forma independiente a la presencia de la CAP y presencia en algunos casos de un segundo ORF corriente arriba del sitio de inicio de la traducción, el cual impide la unión al ORF principal reduciendo de esta forma los niveles de mRNA, y por ende la producción de proteínas [91][57]. El otro tipo de estructura es conocida como `estructura 5 cap` la que formalmente se asocia al extremo 5 del gen tiente una función muy importante en el proceso de traducción ya que funciona como un sitio de unión a factores de iniciación de la traducción y promueve la unión con subunidades ribosomales 40S en conjunto con otras proteínas de preiniciación de la traducción y ayudan a estabilizar el mRNA en los policromas [91][57].

`Intrones en regiones UTR`: las UTR no son exentas de presentar intrones, tiene como característica ser el doble de largos que los intrones inmersos en la región codificante y son característicos de los genes con funciones regulatorias ya que favorecen la ganancia de elementos reguladores de la transcripción. [57].

- `Intron`: los intrones son fragmentos de DNA los cuales están involucrados en el transcrito primario del RNA pero quedan excluidos en el proceso de la traducción por medio del splicing o corte y empalme de exones[3]. Entre las principales funciones que tienen los intrones es ser fuente de RNA no codificante, contener elementos reguladores de la transcripción, contribuir al splicing alternativo, potenciar del cruce meiótico y contener señales para la exportación del mRNA. Otra de sus funciones es tener un papel secundario en los niveles de transcripción ya que se ha establecido una correlación inversa entre la longitud de los intrones y los niveles de transcripción. Los autores mencionan que esto se debe aparentemente a que los genes con altas tasas requieren una rápida respuesta a condiciones cambiantes y los intrones pueden ser perjudiciales para este proceso [91].

- `Región 3 UTR` La región 3 UTR es una secuencia ubicada corriente abajo de la región codificante y sirve como sitio de unión a muchas proteínas reguladoras así como microRNAs, lo que la involucra en numerosos procesos reguladores. Esta región también está encargada de la escisión del transcrito, la poliadenilación, la traducción y la localización del ARNm. En comparación con la región 5 UTR su estructura está sujeta a menos restricciones lo cual la hace propicia a tener un mayor potencial evolutivo, por ejemplo como se ha observado muchos genes pueden tener alternantes regiones de poliadenilación.[91]

2.1.2. Predicción de arquitectura Génica en especies de tunicados y cefalocordados

A continuación se presentarán las principales pipelines existente utilizadas en la anotación genómicas de los tunicados.

Augustus : prediccion de arquitectura Genica en Botryllus schlosseri

.

`Augustus` es un programa basado en modelos HMM, el cual puede predecir tanto splicing alternativo como arquitecturas génicas completas (UTR, exones, intrones). `Augustus` es capaz de incorporar información proveniente de alineamientos tanto de mRNA como de transcritos y EST. Basado en datos de humano se estableció que puede predecir un 77 % de los genes correctamente del genoma humano [107]. Como características de funcionamiento de entrenamiento y generalidades del algoritmo se tiene:

- `Evidencia extrínseca`: se puede entrenar con información parcial de un gen ya conocido dando "pistas" al programa. Estas pistas están asociadas a una tabla de etiquetas biológicas del gen propias de `Augustus` tales como: dirección cadena, marco de lectura, codón de parada o inicio, exones, CDS, UTR, etc. Cada una de estas etiquetas son separadas por la etiqueta biológica establecida, posteriormente son concatenados en un solo archivo y se usan simultáneamente. El soporte de la predicción es evaluado para cada uno de estos grupos [107]

- `Splicing Alternativo`: cuando los alineamiento de los transcritos se contradicen unos con otros, pero los alineamientos generado por `TRASMAP` demuestran una similitud tal que se comprueba que proviene de la misma secuencia, se consideran posibles variables del trascrito y se prosigue a agrupar todos estos hits en un solo grupo al tiempo que se evalúa la similaridad de los CDS construidos a partir de los exones predichos por `EXONPHY` y de esta forma confirmar la pertenencia a un mismo transcrito [107].

  En un segundo paso se evalúa si dos grupos de transcritos previamente establecidos son coherentes si y solo si son compatibles con una arquitectura génica común y es a partir de alineamiento de los grupos compatibles donde se establecen las verdaderas variantes de un mismo transcrito.

  Posteriormente se usa el algoritmo de Viterbi para una última evaluación de los grupos compatibles y para corroborar la presencia de sitios de splicing, terminado en este paso los grupos corroborados como miembros de un mismo transcrito en agrupaciones o en genes. Se evaluado en las etiquetas del paso de Evidencia extrínseca en donde se evalúa que el soporte no sea menor al 86 %.

- `Preprocesamiento de Candidatos`: para la predicción de intrones AUGUSTUS evalúa los posibles sitios de splicing por medio de un consenso de GT/GC-AG y descarta todas las secuencias que tengan un número muy alto de variantes incompatibles y a su vez se encuentren pobremente soportadas (menos del 10 %) son descartadas. [107]

- `Propuestas de ortólogos (TRAN MAP)`: genera alineamientos entre CDS de diferentes especies por medio de `BLAT` el cual solo funciona con transcritos que tengan más de un 95 % de identidad. A esta predicción se le suma la ejecutada por `Blastz`, en la cual se alinean los cDNA contra el genoma. Estos dos procesos le permite a `AUGUSTUS` eliminar los parálogos, ya que se ejecuta un filtro de sintenia, aunque no logra resolver la problemática de las copias en tandem [107].

- `Pseudogenes procesados`: se hacen alineamientos de `BLASTZ` de mRNA contra el genoma y se evalúan características como cantidad de intrones procesados; la ausencia de sitios de empalme conservados y posición y longitud de la cola poli A para poder catalogarlo como un posible pseudogen [107].

SNAP : Prediccion de arquitectura Genica en Oikopleura dioica

`SNAP` o `Semi-HMM-based Nucleic Acid Parser` por sus siglas en inglés, es un predictor de genes *ab initio* en el cual se predicen las arquitecturas genícas a través de HMM. Este programa comparte muchas similitudes con `GENSCAN` ya que ambos basan sus predicciones genicas en la identificación de patrones basados en la probabilidad dada por los HMM de las diferentes estructuras asociadas a la arquitectura del gen. [62]

- `Predicción de intrones`: usa seis modelos de predicción de sitio de corte y empalme para establecer los posibles intrones, con estos modelos evita los codones de parada en los sitios de splicing.

- `Otras Predicciones`: evalúa cada hebra por separado así exista superposición de genes en regiones intronicas [62]

- `Codones Inicio y Parada`: calcula los correspondientes codones de inicio y de parada [62].

`Fgenesh` : Prediccion de arquitectura Genica en Branchiostoma floridae

`Fgenesh` es un predictor de genes *ab initio* el cual predice las arquitecturas genéticas a través de HMM compartiendo muchas similitudes con `GENSCAN` [24], y `Genie` [52],
Las diferentes etapas del análisis son:

- `Predicción de Exones`: genera una lista de los Exones potenciales teniendo en cuenta todos los ORF, usando funciones discriminantes similares a programas previos cómo

`Fexh` (Find exon) y genera cuatro grupos donde los umbrales dependen de la cantidad de GC y de tener el puntaje más alto según el programa.

- `Orden de Exones`: organiza los exones de acuerdo a la proposición final de cada exon 3' [100].
- `Selección de Exones Correctos`: escoge los exones que en contexto con el resto de exones tenga sentido y poseen los más altos puntajes , puntaje basado en funciones discriminantes lineales desarrolladas para exones [100]
- `Agregar promotores`: se agregan los promotores o la Poli A si se predicen con mejores puntajes, puntaje basado en funciones discriminantes lineales desarrolladas para poli A y promotores. [100]

2.1.3. Modelo de prediccion genica en especies sin anotación y ensambladas de de novo: D. vexillum

`GeneID` : Prediccion de arquitectura Genica en D. vexillum

Es un Programa de predicción de genes con una estructura jerárquica especialmente diseñado para genomas de novo.

- `Entrenamiento de GeneID`: entre los archivos iniciales para poder predecir genes por `GeneID` [15] se encuentra un archivo de parámetros el cual es el resultado del entrenamiento del programa, dicho entrenamiento consiste en calcular las matrices de Position Weight Arrays (PWAs)[15]. La escogencia de secuencias usadas para el entrenamiento son básicas para detectar genes.
- `Predicción de Sitios de Splicing y Codones`: se predice sitios de splicing y codones de inicio y de parada y son puntuados de acuerdo a PWAs o a los HMM para los sitios de splicing, codón de inicio y codón de parada para así poder generar un HMM para poder predecir genes en el DNA [15]. Para este paso es necesario un archivo gff de la especie que servirá para entrenar y los archivos fasta correspondientes. Se recomienda que sean más de 500 modelos los usados en el entrenamiento aunque existe la posibilidad de flexibilizar el programa si se tiene menos de 500 modelos. Para esta tesis fueron usados las secuencias de los tunicados *O. dioica* (Oidi) y *C. intestinalis* (Ciin) para entrenar el programa teniendo en cuenta que son las especies de los tunicados para las cuales las anotaciones son las más estables.
- `Predicción de Exones`: los exones se predicen basándose en la suma de los puntajes de los los sitios de splicing y codones de parada que se describieron anteriormente unido a la razon de logaritmo de verosimilitud de un HMM aplicado para detectar lo codificado en el DNA [15]

- `Ensamblaje arquitectura del Gen`: basados en los exones previamente predichos se establece la estructura final del gen, maximizando la suma de los puntajes de los exones [15]

2.2. Metodología

Con el fin de predecir genes en el genoma de la especie *D. vexillum* (Dive), se ejecutó el programa `GeneID V 1.4.4`, entrenado con dos modelos génicos estables existentes para las especies *O. dioica* (Oidi) y *C. intestinalis* (Ciin). El proceso completo de aproximación a genes del SII se observa en la grafica **2-1**. Es importante resaltar que los parámetros usados para la predicción fueron:
geneid -3 -P $\prec archivodeparametros \succ \prec Archivofasta \succ$ -A $\rightarrow$ Archivo de Salida
Se hizo predicción de sitios de inicio, sitio donor sitio, sitio aceptor, primer exon, exón final y una predicción de la secuencia proteica. Todos los resultados se concatenaron en el archivo EspecieGoldenStandard_ciona_fasta.gff3 o su correspondiente anotación cuando se utilizaron las anotaciones de *O. dioica* para entrenar y hacer la anotación dependiente de la información de esta especie. Finalmente del archivo de salida de `GeneID V 1.4.4` se tomaron las secuencias proteicas predichas por el programa generando un archivo multi fasta. Posteriormente sobre cada secuencia de proteínas se detectaron los dominios correspondientes utilizando la información existente para las proteinas asociadas al SII de las especies anotadas por el Ensembl y que fueron curadas en el capitulo anterior usando el golden standard. Los dominios para las especies de tunicados *C. intestinales* y *C. savigny* y otros vertebrados *P. marinus*, *L. chalumnae* y *D. rerio* `TIGRFAMs`, `SUPERFAMILY`, `PIRSF`, `PANTHER`, `CATH-Gene3D` y `PFAM`. Los dominios fueron recuperados para cada sistema de annotación usando la función hmmfecth, en donde Modelos.hmm corresponde al nueva base de dominios extraidos de la base de datos original llamada en nuestro ejemplo como Database.hmm y protein_domain_list corresponde a la lista de dominios de proteinas que fueron anotadas para la especies mencinoadas anteriormente.
hmmfetch -o MODELOS.hmm $\prec Database.hmm \succ \prec protein\_domain\_list \succ$
Posteriormente se construyo el indice para Modelos.hmm y posteriormente se realizó el escaneo de dominios.
hmmscan -o $\prec File.out \succ$ –tblout $\prec output\_tabular.tab \succ$ –domtblout
$\prec output\_tabular\_detallado.tbldom \succ \prec MODELOS.hmm \succ \prec Input.fa \succ$
En donde los parámetros tblout y –domtblout fueron usados como opciones de formato de las salidas y el archivo multifasta de las proteínas detectadas por `GeneID V 1.4.4` como subject o Input.fa. Es importante aclarar que los MODELOS.HMM corresponden a las modelos HMM existentes en cada sistema de anotación de dominos. Posteriormente se generó una salida para la cual se clasificaron los hits como positivos siguiendo los valores estandarizados por cada base de datos. Estas salidas de hits positivos fueron organizadas e interceptadas siguiendo el mismo procedimiento utilizado en el capitulo 1 para las demás especies de

tunicados, es decir, se aplicó la función de reducción y posteriormente se interceptaron los dominios de las arquitecturas de dominios gold standard siguiendo las reglas de Orden, Desorden y Blast (O, D, B). En la figura **2-1** se resume el procedimiento final utilizado. Posteriormente se utilizaron los mismos criterios para la definición de los dominios asociados a los módulos de Señalización, Efector y Reconocimiento del sistema inmune de tunicados siguiendo la clasificación propuesta por [122].

2.3. Resultados

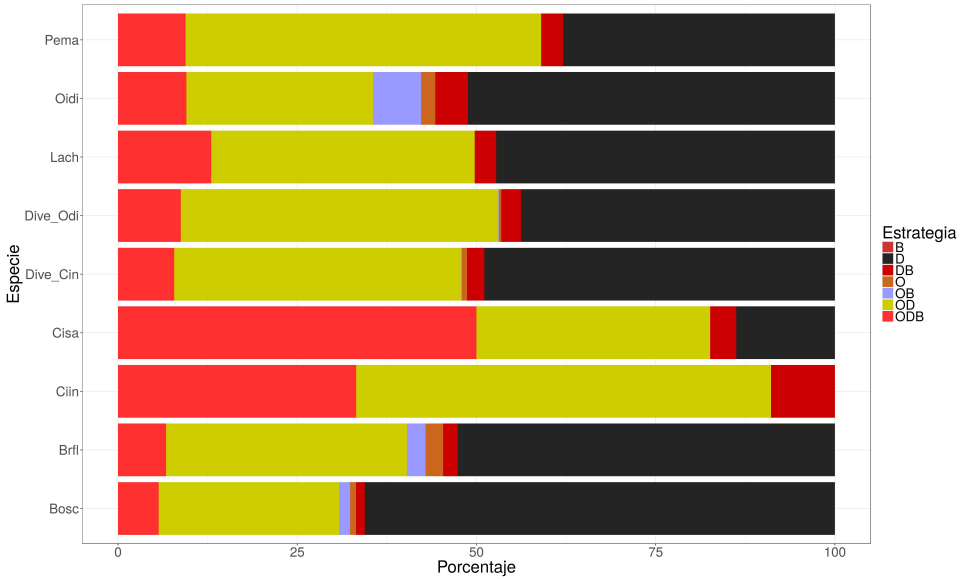

Figura 2-3: Frecuencias de las diferentes estrategias de anotación: **ODB**, con *Nasonia vitripennis* como Gold Standard

Cuando se comparan los modelos con *Nasonia vitripennis* como Gold standard, se evidencia que la estrategia de **OD**, es la más predominante con un incremento en **DB 2-3**. Ahora cuando se compara con *Apis mellifera* Figura **2-4** se evidencia que las estrategias de **OD**, es la más efectiva y que se encuentra igualmente bien representado **DB** y **B** solo, y presenta comportamientos similares a Brfl y Pema como se observó en el capitulo 1.
Al comparar las estructuras Golden Standard de *Homo sapiens* como lo muestra la Figura **2-5**, se observa un predominio de las estructuras Ordenadas y con anotaciones por Homología. Se observa que al igual que con humano, al usar con textsf*Mus musculus* Figura **2-6** como Gold Standard, la estrategias asociadas a **O** se ven favorecidas aunque están seguidas de cerca por la estrategia **D** y no se ve tan representado la estrategia de homología **B** como en Humano.
Ahora, las estrategias **ODB** predominantes en la anotación de las proteínas en Dive, cuando se evalúan con *Acyrthosiphon pisum* como las arquitecturas Gold Standard, son **D**esorden las

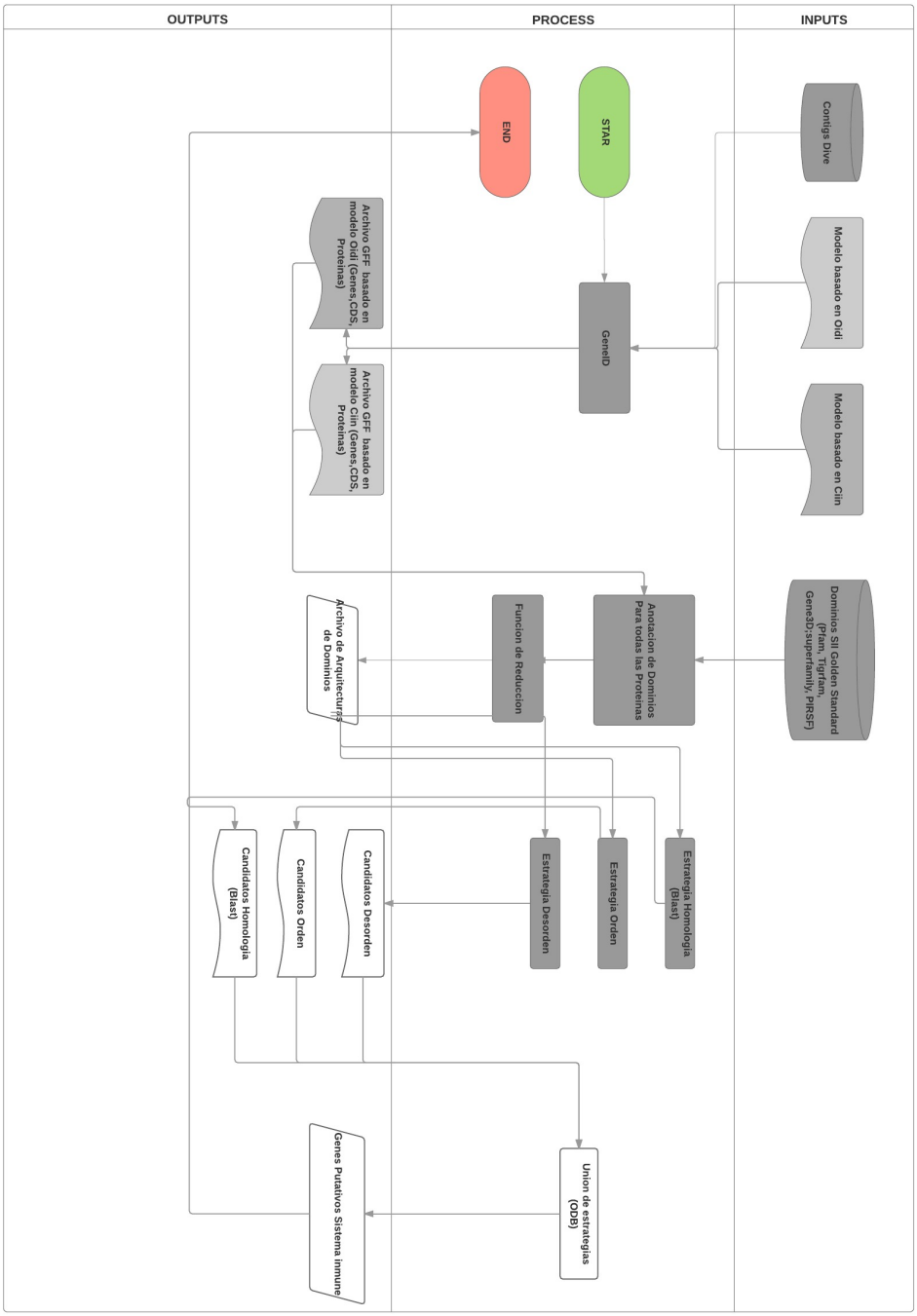

Figura 2-1: Pipeline para obtener las coordenadas y anotación de dominios en la especie *D. vexillum* a partir de dos especies de referencia (Oidi) y (Ciin) e intersección con los dominios gold standard

predominantes, con cierto número de proteínas establecidas por homología **B**.
Con el fin de observar la distribución tanto de genes como de proteínas identificadas por `GeneID` y que posteriormente fueron anotadas por la estrategia ODB como se observa en la

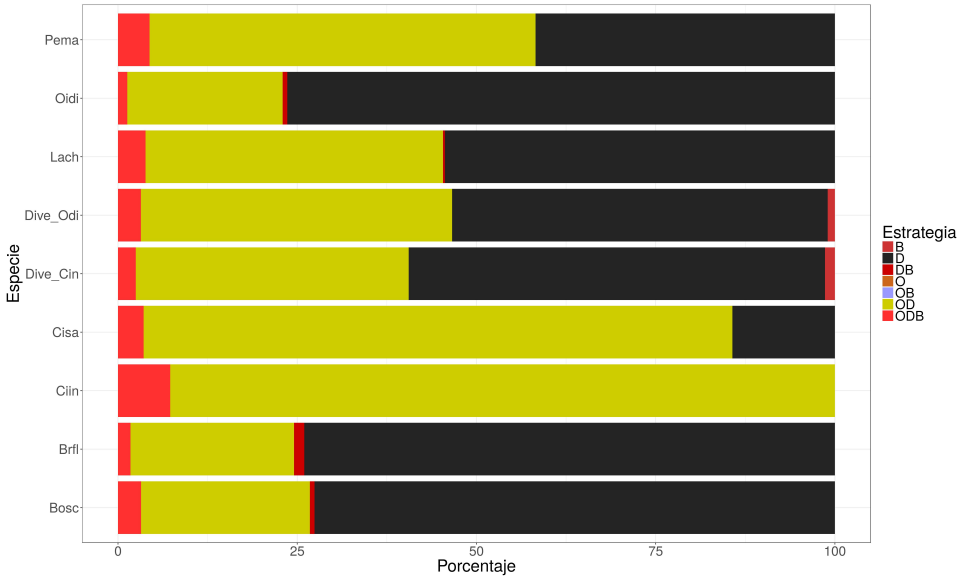

Figura 2-2: Frecuencias de las diferentes estrategias de anotación: **ODB**, con *Drosophila melanogaster* como Gold Standard

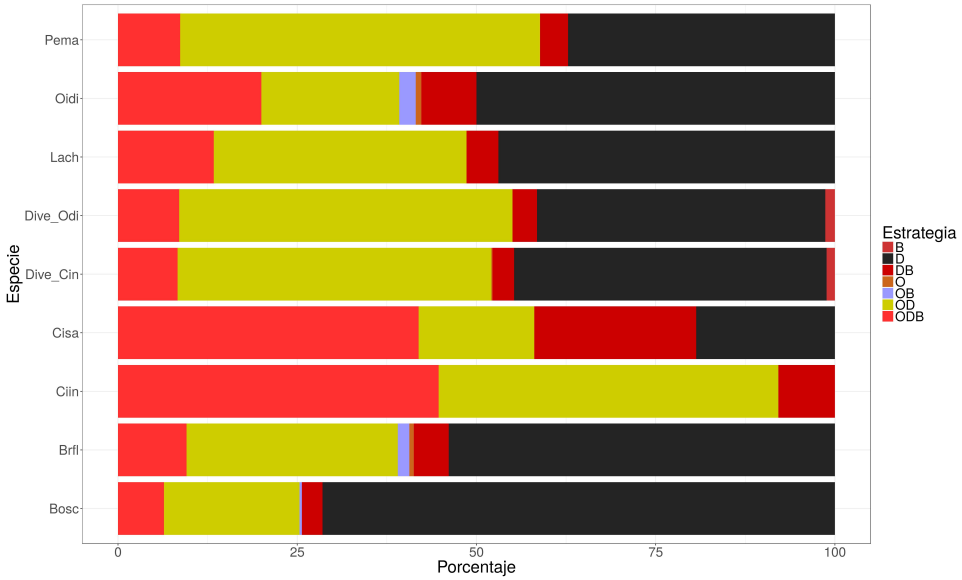

Figura 2-4: Frecuencias de las diferentes estrategias de anotación: **ODB**, con *Apis mellifera* como Gold Standard

tabla **2-1** fue necesario en primer lugar correr de forma separada O, D y B hasta realizar la comparación de las estrategias ODB como se describió en materiales y métodos. Note que en la tabla cada una de las especies de referencia soporta un número diferente de genes. Este tipo de diferencias de anotación depende del modelo escogido. En nuestro caso para las dos especies *O. dioica* y *C. intestinalis* se han reportado dinámicas diferentes de evolución que

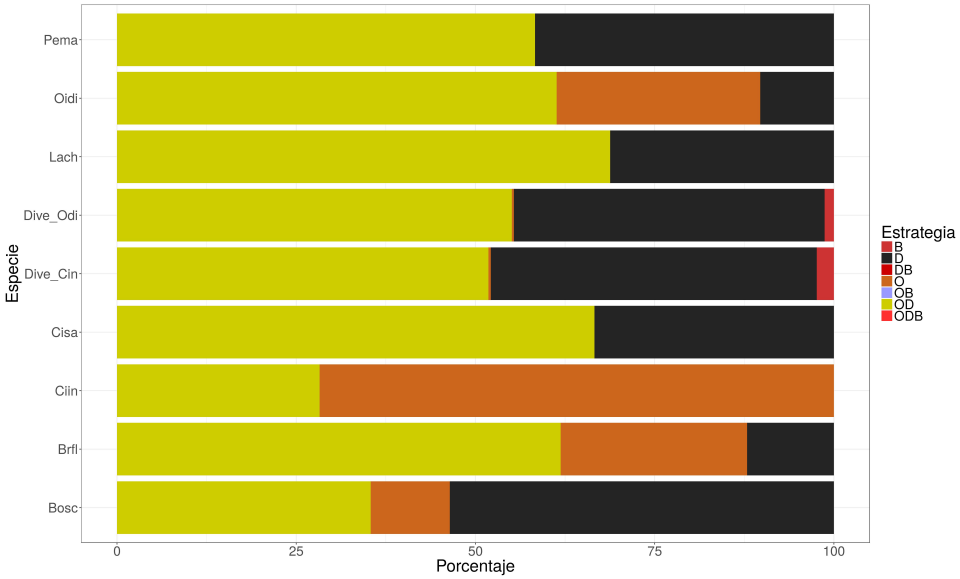

Figura 2-5: Frecuencias de las diferentes estrategias de anotación: **ODB**, con *Homo sapiens* como Gold Standard

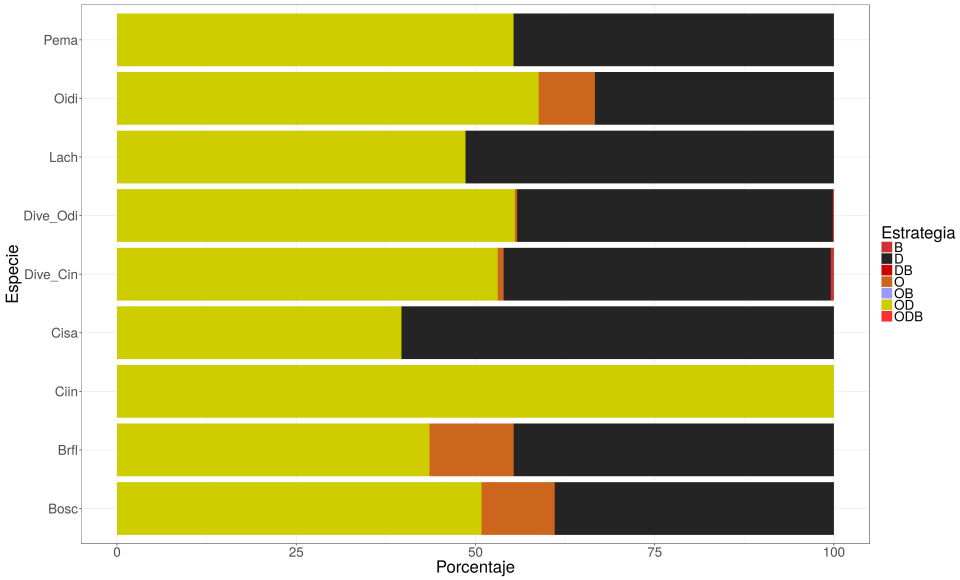

Figura 2-6: Frecuencias de las diferentes estrategias de anotación: **ODB**, con *Mus musculus* como Gold Standard

conllevaron por un lado a la reducción del genoma y el número de genes como en *O. dioica* y por otro a la perdida de homologada de algunos genes de *C. intestinalis* con otros cordados como se mencionó en la introducción del trabajo. Adicionalmente se observa que el número de secuencias que se detectan por homología de Blast fue similar cuando se usaron ambos modelos de referencia para la anotación de genes en *D. vexillum*.

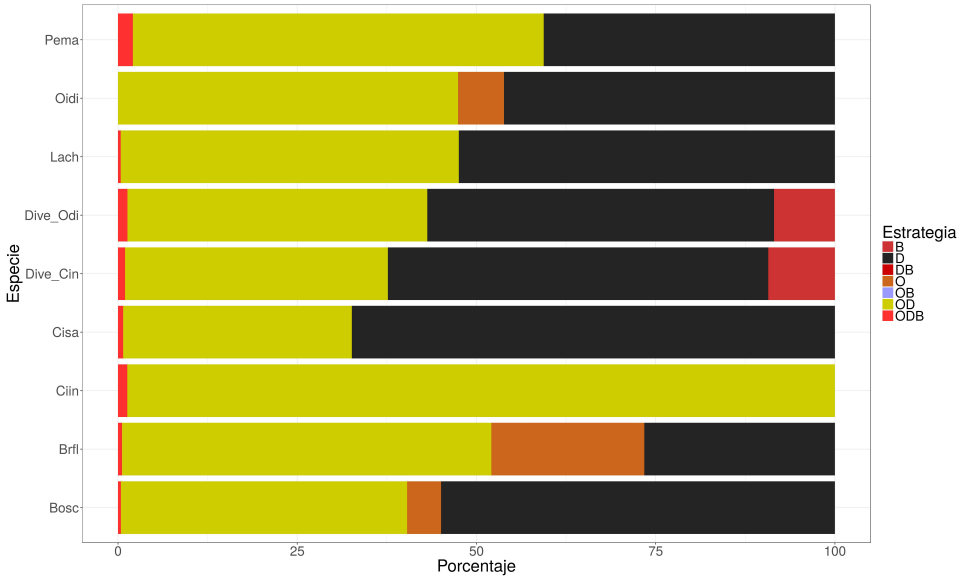

Figura 2-7: Frecuencias de las diferentes estrategias de anotación: **ODB**, con *Acyrthosiphon pisum* como Gold Standard

Con el fin de visualizar cuáles son los dominios más predominantes en cada uno de los módulos definidos por [122] en tunicados y observar si hay diferencias aparentes entre los dos modelos usados para la anotación de genes se observó cual era el porcentaje que representaba cada uno de los dominios hallados por cada módulo. Se observa en la Figura **2-7**, que el modulo que más se diversifica es el de señalización seguido por el de reconocimiento. En el primero el dominio predomiante es **Ank** mientras que en reconocimiento se encuentra las **Pkinasas** y el dominio **fn**. Por el contrario se observa un claro dominio del dominio **Efector** en el módulo efector. Se encuentra que no hay mayor diferencia entre los módulos propuestos por los dos modelos de referencia.

Tabla 2-1: Distribución final del número de proteínas putativas asociadas al sistema inmune en *D. vexillum*

Referencia	Genes Anotados	Proteinas Anotadas	Proteinas Orden	Proteinas Desorden	Proteinas con homologia	Total
C. instestinalis	8350	8350	104	193	113	234
O. dioica	3162	3162	96	184	91	210

2.3.1. Módulos del Sistema Inmune

Ahora, con el fin de identificar cuales son las diferencias en la cantidad de dominios que se logran identificar por medio de los dos modelos de referencia para la predicción de genes, se decidió comparar el número total de dominios asociados a proteínas putativas establecidas por ODB en la especie *D. vexillum*, tanto para los genes predichos con Dive-Ciin como Dive-Oidi. Como lo muestra la figura **2-9**, en la que se observa que las predicciones por ambos modelos predicen de forma satisfactoria 105 dominios comunes, mientras que dominios predichos en Dive Ciin y no en Dive Oidi fueron 85 y en el caso contrario solo dos dominios no pudieron ser hallados por el modelo Dive-Ciin.
Al evaluar cuales de los dominios asociados al módulo efector fueron predichos en Dive basados en los dos modelos de referencia, se encontraron 5 dominios asociados a este módulo y que la totalidad de los dominios predichos en Dive-Oidi fueron también controlados en Dive-Ciin, pero que al comparar en sentido contrario se observa que el modelo de Dive-Oidi no logra identificar el dominio **WD40** (PF00400) como lo muestra la gráfica **2-10**.
Se logró identificar que los dominios encontrados por ambos modelos son para el Modulo de señalizacion fueron 11 y que el modelo Dive-Ciin fue capaz de identificar dos dominios más que el modelo Dive-Oidi, dichos dominios son: PF00090 (**TSP** y PF00400 (**W40**) como lo muestra la gráfica **2-11**
En los dominios asociados al módulo de reconocimiento se identificaron 9 dominios identificados por ambos modelos, y tan solo un dominio identificado solamente en Dive-Ciin PF00400 el dominio **W40 2-12**.

2.4. Discusión

De la tabla **2-1** se deduce que existe una diferencia marcada en el número total de genes predichos por `GeneID` basados en los dos modelos de referencia basados en las especies *O. dioica* y *C. intestinalis*, siendo el modelo Dive-Ciin más robusto que su contraparte Dive-Oidi ya que se duplica en el primero las predicciones con relación al segundo. Sorprendentemente al momento de evaluar las predicciones de genes del sistema inmune la diferencia es casi imperceptible pero existente. Se podría decir que debido a la alta reducción que presenta el genoma de Oidi sumado a la reducción tanto en el número como en la longitud de los intrones, que puede ser dado por el tiempo tan corto de vida que presenta. Algunos de estos fenómenos evolutivos marcan el tipo de genes y las arquitecturas que pudieron ser conservadas en cada una de las especies y estos son en conjunto factores decisivos que soportan que el modelo Dive-Ciin fuera más efectivo, al ser esta especie ma¿ás cercana tanto al tiempo de vida como a la cantidad y longitud de intrones con relación a la especie *D. vexillum*, por lo que resulta en una mejor predicciónde genes sobre la especie.
De igual forma se observó que la reducción en el número de predicciones de genes influenció la identificación de cierto dominios, dejando en evidencia nuevamente que el modelo Dive-

Ciin logra mayores aciertos, aunque al comparar la totalidad de los dominios se observa que existen dominios aunque pocos que pudieron ser identificables únicamente en el modelo Dive-Oidi, lo que sugiere que esta última puede llegar a ser en cierto punto un complemento a la anotación de las predicciones logradas por el modelo Dive-Ciin.
Además se pudo evidenciar que los modelos Dive-CIin y Dive-Oidi no presentan mayor diferencia al momento de comprar el número de dominios pertenecientes a los diferentes módulos del sistema inmune, también se observa que el módulo efector está claramente dominado por los dominios efectores, mientras que por el contrario el dominio de señalización posee más diversificación de dominios, seguido de cerca por el modulo reconocimiento, lo cual era de esperarse ya que la señalización y el reconocimiento en una especie colonial es de vital importancia, por lo que se puede observar una disminución de diversidad en dominios efectos.
Se encontró que el grupo de los Hymenoptera, compuesto en nuestro estudio por *Nasonia vitripennis* y *Apis mellifera*, aumentaron su aporte a las arquitecturas sumado a que las estrategias que involucraban homologías de Blast aumentaron en comparación con el grupo Díptera compuesto por *Drosophila melanogaster* . Las arquitecturas Golden de las proteínas asociadas al sistema inmune en *Drosophila melanogaster* no se encontraran en la estrategia de Orden sino que se favorecieran las estrategias basadas en Desorden. Mientras que el grupo de crustáceos que representa los icetos *Acyrthosiphon pisum* las estrategias de Desorden son predominantes. Cuando se evalúan las estrategias en mamíferos, se denota que las estrategias de Orden se conservan, lo que puede indicar que después de la diversificación del sistema inmune adaptativo, las arquitecturas provenientes de Tunicados se mantuvieron constantes en los vertebrados sin grandes modificaciones, pero al tener una baja homología de secuencia completa se puede deducir que los cambios se dieron en cambio de orden de dominios o de incremento de sustituciones no sinónimas.

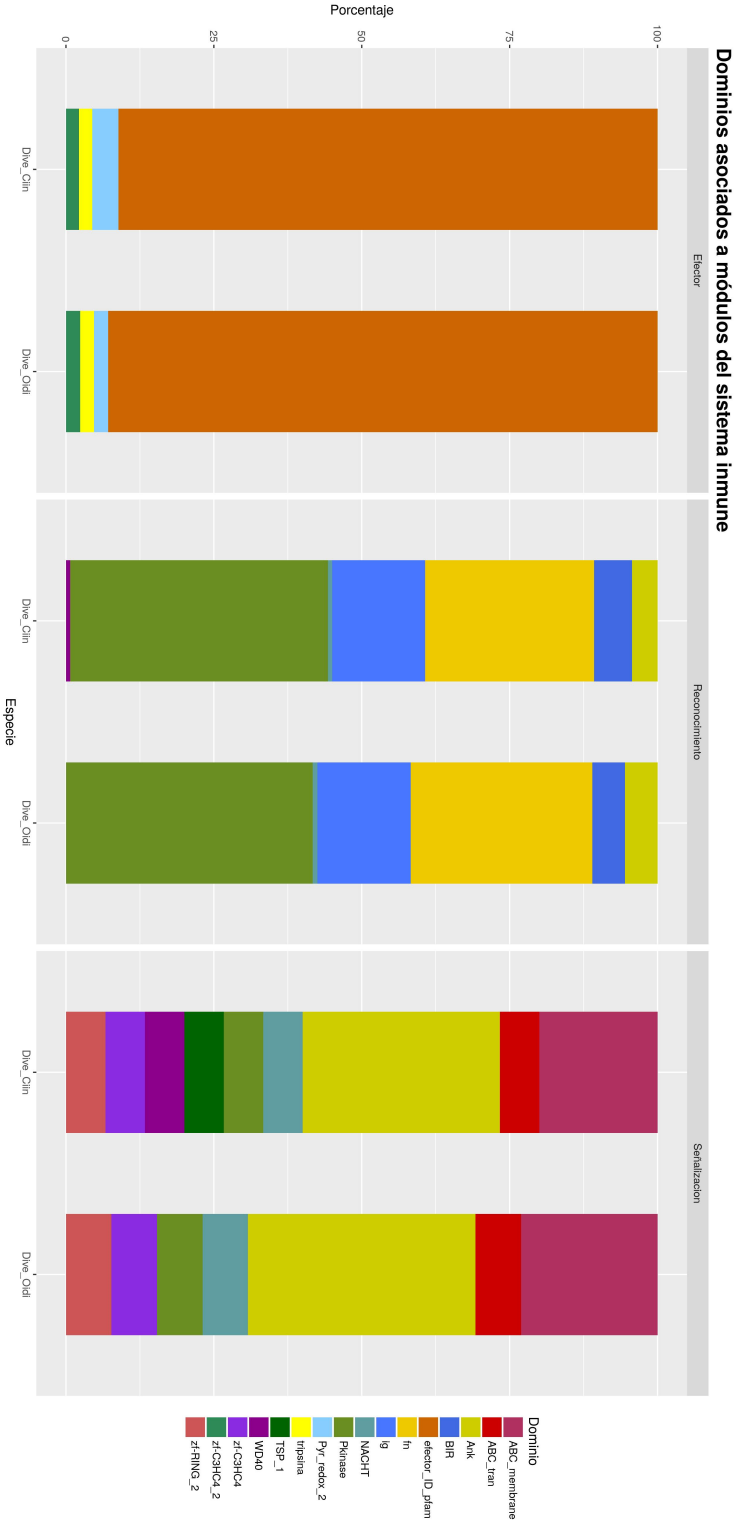

Figura 2-8: Comparación de los porcentaje de los diferentes dominios asociados a los tres módulos del sistema inmune en los modelos de genes en Dive producidos por genID basados en modelos de Oidi (Dive_Oidi) y Ciin (Dive_Ciin)

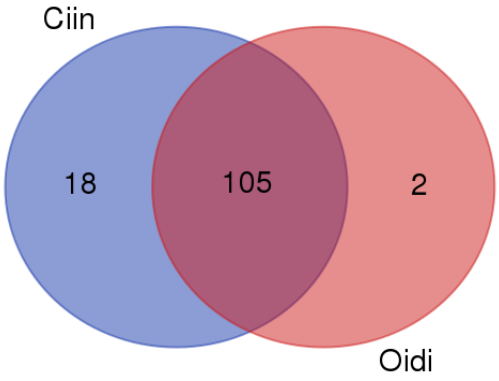

Figura 2-9: Diagrama de Venn que muestra el número Total de Dominios de todos los módulos asociados a proteínas putativas predichas por ODB en *D. vexillum* Dive usando las dos referencias

Figura 2-10: Diagrama de Venn que muestra el número Total de Dominios del modulo Efector asociados a proteinas putativa predichas por ODB en *D. vexillum* Dive usando las dos referencias

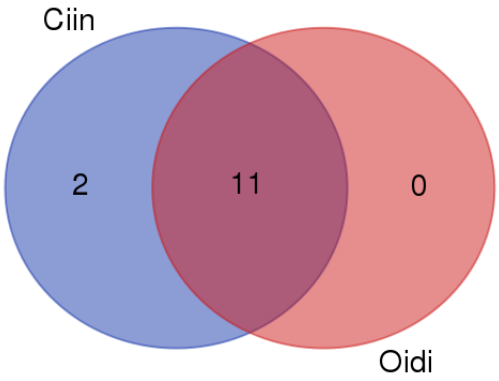

Figura 2-11: Diagrama de Venn que muestra el número Total de Dominios del modulo Señalización asociados a proteinas putativa predichas por ODB en *D. vexillum* Dive usando las dos referencias

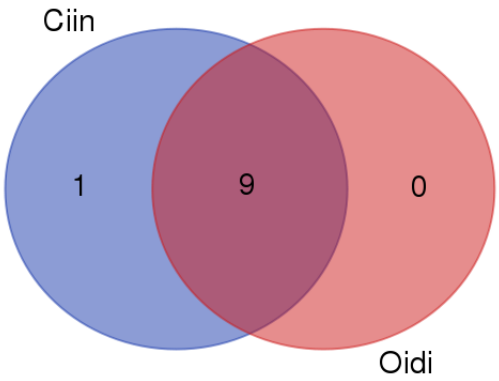

Figura 2-12: Diagrama de Venn que muestra el número Total de Dominios del modulo Reconocimiento asociados a proteinas putativa predichas por ODB en *D. vexillum* Dive usando las dos referencias

3 Modelo explicativo de la evolución de dominios asociados al sistema inmune como propuesta para explicar las dinámicas del sistema inmune

3.1. Introducción

3.1.1. Principios básicos de modelos de ganancia y perdida de genes

El estudio evolutivo de ganancia y perdida de genes desde la escala genómica ha tomado relevancia en genómica comparativa en los últimos años. Muchos de los estudios actuales derivan de los estudios pioneros sobre la evolución de la super familia de las globinas. De esos estudios, a mediados de los 90, se había acuñado la idea de que los miembros de una misma familia evolucionaban de manera concertada, que Nei [82] resume asi "cada uno de los miembros de una familia génica evolucionan al mismo ritmo o como diríamos en un sentido musical al unísono". En este mismo trabajo Nei replantea que el modelo de evolución concertada no es el único mecanismo que puede explicar la evolución de las familias génicas e incluso replantea en su tiempo que era innecesario invocar siempre al proceso de conversión génica como el mecanismo que varios autores hasta antes de los 90 utilizaban para explicar el alto polimorfismo por ejemplo en la familia de genes del MHC por su traducción del inglés (Complejo Mayor de Histocompatibilidad).

De varios trabajos realizados por Nei sobre los genes del MCH, se planteó un nuevo concepto sobre la ganancia o pérdida de genes conocido como el "Modelo de nacimiento y muerte de genes". De allí surgió la discusión de si la evolución concertada era el mejor modelo que explica la evolución de todas las familias génicas. Dentro de los mejores ejemplos estudiados se tiene el de la evolución de la familia de genes Wingless (wnt) en metazoarios [2] y de otros ejemplos como los genes de choque térmico (hsp70), genes asociados con el desarrollo, histonas, amilasas, genes de las ubiquitinas y asociados al sistema sensor como quimio receptores, receptores olfativos. Interesantemente, los modelos de nacimiento y muerte de genes ha sido explicativo no solo para los genes del MHC sino también para otros genes asociados al sistema inmune innato como de algunas inmunoglobulinas ancestrales y de familias de genes asociados al SII [82]. Un tema adicional planteado Cañestro et al, [2] es la discusión

sobre si la perdida de genes es relevante para establecer la divergencia de especies dentro de un phylum, es decir si la perdida de genes puede considerarse neutra o causal o positiva en los procesos de formación de nuevas especies.

Aunque el tema de evolución de familias de genes aparentemente parece muy antiguo, sólo a partir de la automatización e incremento acelerado de genomas ensamblados de especies de grupos taxonómicos que permitan hacer comparaciones consistentes, este campo ha tomado relevancia y ha ganado terreno en la genómica comparativa a gran escala. En general los métodos fundamentan su metodología de cálculo son basados inicialmente en las predicciones de homología y ortología. Los métodos de ortología a su vez están basadas en tres aproximaciones principales: los métodos dependientes de árboles, los métodos dependientes de grafos y los métodos híbridos que combinan las dos aproximaciones [65].

Un modelo aplicado para cuantificar ganancia y perdida de genes en el Programa Count

Count es un programa diseñado para el análisis de perfiles numéricos en una filogenia dada, es decir las relaciones filogenéticas de las especies deben ser conocidas *a priori*. Su función principal es analizar perfiles derivados de una filogenia proveniente de genes homólogos. El programa provee una reconstrucción ancestral dada una topología del árbol permitiendo inferir características específicas de ganancias o pérdidas e incluso de duplicaciones a lo largo de la filogenia dada. Este programa implementa los modelos estadísticos de parsimonia de Dollo y Wagner, haciendo un énfasis en la cuantificación de presencia y ausencia de genes, así como los métodos probabilísticos que implican un modelo filogenético de nacimiento y muerte de genes [31].
En esta tesis analizaremos la perdida de dominios basado en el modelo de Dollo aunque su aplicación más conocida es su aplicación a las estudio de genes. Este modelo está basado en el precepto de que si un gen ha sido perdido en un determinado linaje no puede ser recuperado por este de nuevo (es decir no puede retornar al estado inicial), sino que debe encontrar caminos alternos (es decir condiciones) para la aparición de múltiples ganancias del mismo gen ya que la combinación fortuita y aleatoria de mutaciones que permita el resurgimiento de la misma secuencia funcional es poco probable (adquirirlo en paralelo o convergencia), en otras palabras es más probable que aparezca una pérdida del gen a que se de origen independiente. Aplicar este concepto permite una simplificación de los análisis evolutivos reduciendo la ambigüedad en la reconstrucción de la historia evolutiva del mismo [42].
En la forma más simple de usos del programa los estados complejos derivados son representados con 1 y 0 representa el estado ancestral primitivo. Para representar el estado se puede formalmente usar cadenas de unos para la presencia del carácter y ceros para la ausencia del mismo, y es posible construir un árbol de parsimonia de Dollo basado en una matriz de presencia ausencia en donde el modelo intenta minimizar el número de reversiones de 1 a 0

y excluye las convergencias[42].
Los valores reconstruidos en los componentes del árbol son basados en la matriz de presencia-ausencia y fuertemente dependientes la topología del árbol. En las ramas del árbol se le van a mapear los diferentes tipos de eventos evolutivos y a su vez se reconstruyen los estados de los caracteres en los nodos internos del árbol[42]. El modelo de Dollo no se debe asumir perfecto por defecto, depende fuertemente de la confiabilidad de la topología del árbol, y debe tenerse atención si los datos violan la irreversibilidad genética-filogenética al producir ganancias de carácter homoplásico Se debe tener en cuenta que no se debe usar para el análisis de genomas procariotas sino en ese caso debe usarse una parsimonia ponderada[42].

3.1.2. Principios básicos de identificación de ortología

La ortología en su definición mas general se entiende como la homología de subregiones genómicas (como los son los genes) compartida entre especies derivadas de un ancestro común y que surgen después de un proceso de especiación en vez de un proceso de duplicación en el genoma de la misma especie que se entiende como paralogia. Debido a esta relación de identidad la función biológica entre estas secuencias está dada por sentada entre los genes ortólogos o por lo menos se espera que parte de la funcionalidad ancestral se mantenga, siendo esto último muy útil en los estudios de genómica comparativa y anotación del genoma, de ahí la importancia de establecer correctamente las relaciones de ortología [68] [64]. Existen dos formas clásicas de aproximarse a la ortologia de genes la primera mediante grafos, el cual genera agrupaciones de secuencias basado en la similaridad de las mismas y dos método basado en construcción de árboles, el cual no solo agrupa sino que también reconcilia con el árbol de secuencias proteicas con el árbol de especies [114]. Aunque existen diferentes aproximaciones para la detección de ortólogos adicional a los métodos dependientes de árboles y métodos basados en grafos también existen métodos híbridos que combinan las dos aproximaciones.

Detección de ortólogos por métodos dependientes de árboles

Estas predicciones principalmente son basadas en métodos dependientes de árboles inferidos para familias génicas. En estos métodos, primero se seleccionan las secuencias homologas entre las especies de estudio y se procede a construir un árbol filogenético a partir de alineamientos. Posteriormente se procede a analizar si el árbol agrupa las secuencias de acuerdo con relaciones filogenéticas consistentes entre las especies. Debido a que los árboles de las especies no necesariamente reflejan las historias evolutivas de los genes se complementan con métodos de reconciliación de árboles. Dentro de los trabajo pioneros se tienen los trabajos evolutivos sobre globinas[49] y el planteamiento del problema de conciliar árboles de genes y los árboles de especies [87]. En el 2005 se presenta una aproximación para la identificación automática de ortólogos y parálogos en genes homólogos[36] y [53]. Cuando se carece de los árboles de las especies dos métodos se han implementado para distinguir entre ortólogos y parálogos unos lla-

mados agrupamientos basados en coeficientes de correlación o Çorrelation Coefficient-based Clustering (COCO-CL; http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/COCOCL/) y los basados en niveles de ortología desde árboles (Levels of Orthology From Trees) (LOFT; http://www.cmbi.ru.nl/LOFT/).

Detección de ortólogos por métodos dependientes de grafos

Los metódos independientes de árboles [65] se construyeron para hacer inferencia de ortología desde los proteomas completos de las especies y aunque no se basan en la idea de reconstruir un árbol desde alineamientos múltiples de secuencias, si fundamentan sus principios en la idea de construir agrupamientos o clusteres de secuencias proteicas homólogas. Por ejemplo construyen las similitudes a partir de todas comparaciones pareadas que pueden basarse en métodos de alineamiento de secuencias heurísticos como BLAST o en los métodos de alineamientos pareados no heurísticos como los de Smith Waterman. La idea principal de estos métodos es recuperar los mejores hits recíprocos o "reciprocal best hits (RBHs)", que son conocidos en otras aproximaciones como BBH o (Bi-directional Best Hits). La idea biológica que se encuentra detrás de estos métodos se basa en el principio de que genes similares deben tener funciones similares, es decir los miembros de un par BBH, son similares en secuencia y en función [41]. El objetivo principal es la reconstrucción de grupos ortólogos basados en un sistema de puntuación. Se pueden distinguir métodos basados en operaciones que buscan el vecino mas cercano como primer filtro para la clasificación de ortólogos putativos entre las mejores hits de las comparaciones pareadas de las especies y la búsqueda de agrupamientos o clusters de grupos ortólogos del acrónimo en inglés "Clusters of Orthologous Groups (COGs) of proteins"[66] . Aproximaciones como InParanoid[96] en el cual el algoritmo permite una eficiente identificación entre ortólogos y out-paralogos. En la primer versión del programa los autores resaltan que el método automáticamente permite discriminar entre los paralogos (out-paralogos) que predatan la separación de las especies que generalmente generan confusión para la asignación de ortólogos verdaderos, lo cuales son confirmados por la identificación de in-parálogos que surgen después de que las especies se separan. En general la heurística de programa se basa en encontrar los conjuntos de pares de proteínas que poseen el score de valor más alto y de tipo simétrico de BLAST que es usado como una semilla para encontrar todos los in-parálogos para cada especie [66], el programa sólo permite hacer comparaciones pareadas entre dos especies, sin embargo MultiParanoid `http://www.sbc.su.se/bandale/multiparanoid/html/index.html` en contraste permite hacer todas las comparaciones para la construcción de grupos ortólogos de múltiples especies. The Ortholuge program `http://www.pathogenomics.ca/ortholuge` se fundamenta en la idea de Inparanoid pero en contraste no usa el score basado en BLAST sino un score basado en razones de distancia filogenética entre los pares de comparaciones e incluyen la idea de eventos de pérdidas de genes [66]. Por otro lado pipelines como OrthoMCL y OrthoMCL–DB integran un algoritmo de Cluster tipo Markov (MCL) para la construcción de los

grupos ortólogos[70]. Los autores describen la aproximación como un proceso de dos pasos: la construcción de un grafo que representa relaciones entre secuencias que posteriormente son subdivididas en subgrafos usando el algoritmo MCL. En el primer paso es importante conocer que tipo de secuencias pueden ser incluidas, como las secuencias nodo son conectadas y como se deben ponderar las aristas entre los nodos para cuantificar las relaciones entre las secuencias. En el segundo paso, el algoritmo de clusterización MCL funciona para la identificación de grupos coherentes usando un índice conocido como "cluster granularity—the inflation index"[70].

- ProteinOrtho: ProteinOrtho es un programa diseñado para poder evaluar relaciones de ortología a partir de la teoría de grafos. Dicho programa ubica los genes en los vértices y las aristas son dirigidas y ponderadas por un peso dado tanto por el bit score como por el e-value de un Blast recíproco entre las dos secuencias evaluadas. Así se define que las aristas representan relaciones de ortología entre ellas. A partir de la construcción del grafo, se puede establecer dos estructuras, el complemento de un grafo, que para el efecto práctico de ProteinOrtho representaría las relaciones de paralogía y el grafo inducido. El complemento de un grafo se define sobre el mismo conjunto de vértices y el establecimiento de las aristas faltantes sin tener en cuenta las aristas del grafo original, mientras que un grafo inducido es un subconjunto de vértices del grafo original, en donde las aristas que conectan esos vértices provienen del grafo original. Para poder establecer relaciones de ortología es necesario tener un grafo inducido que sea cografo [68]. Esta alternativa surge como una opción que puede ser complementaria a los cálculos de ortología que usan árboles filogenéticos, que aunque presentan de forma gráfica y fácilmente entendible los procesos de paralogia y ortología, son computacionalmente costosos. En muchos casos tienden a fallar cuando se evaluándote las familias proteicas altamente complejas o cuando se comparan un gran número de especies [114]. Además, ya que los modelos basados en árboles presentan ambigüedad en la comparación de secuencias generadas por la presencia de genes parálogos, el método ProteinOrtho puede soportar la correcta predicción de ortólogos. Entonces entre otras aplicaciones esta aproximación puede resolver la ortología cuando existan dos o más genes en un linaje, que son colectivamente ortólogos entre uno o más genes en otro linaje si el evento de duplicación es posterior al evento de especiación lo cual conllevara al programa a la identificación de coortologos[68].

Detección de ortólogos por métodos híbridos

Estas aproximaciones combinan métodos basados en árboles por su importancia en la resolución que tienen a nivel filogenético y los métodos basados en grafos que incrementan las escalabilidad requerida para los estudios amplios del genoma. Por ejemplo, los métodos utilizados por TreeFam [99] se basan en la definición de familia de genes como *"un grupo de genes que descienden desde un gen sencillo que se encuentra presente en el ancestro*

común más cercano y que es escalable hipotéticamente al ancestro de todos los animales". Este método se basa en el uso de longitudes completas de los genes y no en sus dominios. Por jemplo, TreeFam utiliza en su aproximación dos tipos de bases de datos: una generada por su flujo automatizado de procesos o llamado TreeFam-B y una segunda parte de árboles curados manualmente o llamados TreeFam-A.
La primera parte del proceso o construcción automatizada de los árboles TreeFam-B se compone de tres pasos principales que se resumen de [99] como sigue:

- Construcción de familias semillas: construcción de (PhIGs) de acrónimo del inglés (Phylogenetically Inferred Groups), que son en si una base de datos generada automáticamente a partir de las familias génicas que descienden desde un gen ancestral común. Estos PhIGs son extraídos de bases datos de hongos y son los utilizados como las semillas para la re-construcción de árboles TreeFam-B. Los grupos son inferidos a partir de previas relaciones filogenéticas de las especies, es decir en un nodo de un árbol de especies se forman clusteres si los genes de dos taxones hermanos son más similares (en distancia de secuencias de proteinas) que éstos con relación a un grupo externo.

- Expansión de las familias semilla a familias completas de otras especies: principalmente se hace por búsquedas de homología en las bases de datos de secuencias de proteínas de otros animales, con una búsqueda rápida de BLAST para detectar genes candidatos a genes homólogos. Luego se utilizan los alineamientos de las secuencias semillas (usando MUSCLE) como entrada para HMMER para hacer la selección de los mejores candidatos de BLAST como genes homólogos. En algunos casos algunos genes pueden quedar agrupados en más de una familia, pero en el proceso de curado manual posterior resuelve estos problemas.

- Construcción de árboles filogenéticos usando familias completas TreeFam-B: se realinean las secuencias de las proteínas de las familias completas de nuevo con MUSCLE y se filtran aquella secuencias muy conservadas usando CLUSTALX con el sistema de score de BLOSUM62 con el fin de calcular el score por columna. Posteriormente este nuevo alineamiento filtrado se usa como entrada del algoritmo NJ del acrónimo del inglés (Neighbor-Joining Algorithm) para la construcción del árbol filogenético.

Otras etapas incluyen la curación manual o construcción de árboles TreeFam-A en que los editores utilizán visualizadores para mostrar y editar los alineamientos como Jalview alignment. Si se encuentran datos inconsistentes se vuelven a correr los programas de BLAST y HMMER si se sospecha que un gen no ha sido incorporado.

Otros métodos usados previamente como en PHOG, por el acrónimo del inglés (Phylogenetic Orthologous Groups), [77] utilizan métodos automatizados para encontrar clusteres de grupos ortólogos en nodos de una árbol taxonómico dado. Como resultado de su método

se genera un árbol de grupos de genes ortólogos. El procedimiento inicia con la construcción de un supergen desde un alineamiento múltiple PHOG; es decir construidas secuencias consenso de los PHOG en cada especie, se corre una estrategia pareada de comparaciones entre pares de especies basada en BLAST PHOG-BLAST. A partir de los BBH o del inglés (bi-directional best hits) se forma un grafo con vértices que representan los supergenes y con BBHs que representan las aristas. A partir de estos grafos se encuentran los componentes conexos del grafo que tienen los mayores BBH y se usa un ordenamiento para iniciar las comparaciones para incluir los nuevos ortólogos al grupo. Éstas mejores relaciones se utilizan como semillas para re-iniciar un proceso de búsqueda de nuevos de ortólogos no incluidos usando alineamientos pareados de posibles candidatos a los supergenes de un grupo ortólogo ya sea por el algoritmo de alineamiento de Smith-Waterman en una primera etapa o con un refinamiento posterior usando el algoritmo de Needleman-Wunsch [77].
Aplicaciones como MetaPhOrs [93] combinan igualmente información de árboles filogenéticos como los construidos por TreeFam y otros repositorios como PhylomeDB, EggNOG, COG, Fungak Orthogroups y OrthoMCL. Para estos repositorios se hacen recomputo de árboles cuando es necesario para la aplicación de su algoritmo. Se hace recomputo de los árboles utilizando Maximun Likelihood para cada familia de genes de EggNOG. Para los datos derivados de OrthoMCL y COG los grupos de se reconstruyeron usando alineamiento con MUSCLE, corte de regiones ricas en gaps y uso de modelos evolutivos implementados en PhyML. Posteriormente por un escaneo de las topologias de los árboles de todas las filogenias de los genes se busca si existen procesos de especiación o de duplicaciones en los nodos internos de los árboles usando el algoritmo de sobrelapamiento de especies implementado en ETE [**?**]. Dependiendo de si las ramas comparten una o mas especies entonces se identifican duplicaciones, pero para nodos en los que sus ramas hijas no comparten especies se establecen entonces las ortologias. Finalmente un score llamado de consistencia que combina los asignamiento de ortologías y paralogías se construye si varios árboles confirman las relaciones de ortología.
Finalmente una de las pipelines mas utilizadas actualmente para el análisis de animales cordados se llama Ensembl Compara [54], principalmente sus recursos permiten computar alineamientos genómicos pareados y múltiples a partir de los cuales sintenia a gran escala, valores de conservación y elementos conservados son computados.

- Familias de genes en Compara: con el fin de detectar de mejor forma la homología y la ortología entre las proteínas ya predichas en conjunto con las proteínas de UniProtKB, se catalogan las proteínas identificadas en Ensembl. Dicho proceso de selección comienza en el momento que se cargan ambos set de proteínas, tanto de Uniprot como por parte de Ensembl, de las cuales se seleccionaron todas las proteínas proveniente de todos los genomas de la base de datos, con el fin de abarcar todo el árbol de la vida (hongos, protistas, plantas, metazoos, cordados, etc.). Dichas proteínas se les pasarán los HMM de genes de la biblioteca de Treefamy de la base de datos de Phanter para clasificarlos en familias ya preestablecidas.

Posterior a la construcción de las familias basadas en Phanter, los curadores del Ensembl detectaron que los clusters formados solo por las familias génicas son de tamaño considerable, y decidieron usar las subfamilias de la misma base de datos. Estos sets finales son alineados y se les asigna una descripción consenso a la familia proteica a la que quedaron asignados. La pipeline tiene como parámetros para considerar si una proteica no pertenece a una familia proteica cuando 1) se cubre menos del 40 % del modelo y es allí donde es clasificada como 'AMBIGUOUS' y 2) si la cobertura es igual a cero es clasificada como 'UNKNOWN' [92], [1].

3.2. Metodología

3.2.1. Parsimonia de Dollo

Para poder evaluar las perdidas y las ganancias de dominios del sistema inmune se usó el programa `Count` [31] bajo el principio de parsimonia de Dollo. Para que el programa funcione se requiere la construcción de dos tipos de archivos de entrada: primero un árbol filogenético en el formato Newick y por otro una matriz con los conteos de dominios por cada módulo del sistema inmune de las especies de estudio. El procedimiento inicia con la carga del árbol filogenético asociado a los tunicados y otros cordados, y posteriormente se carga la Matriz. Posteriormente se escoge el modelo de "Dollo parsimony"para realizar el computo y cálculo de ganancia y perdida de dominios.

3.2.2. ProteinOrtho

Basados en la lista de genes asociados al sistema inmune calculados por la estrategia ODB, se obtuvieron la lista de proteínas putativas al sistema inmune para cada una de las especies y por cada uno de los módulos asociados a los tunicados. Se construyeron archivos multifasta independientes por cada especie, correctamente etiquetados. Posteriormente se corrió el programa proteinOrtho con los siguientes parámetros:

- proteinortho5.pl -graph -verbose -project=NombreProyecto $\prec archivomultiFasta \succ$

3.3. Resultados

3.3.1. Ganancia y Perdida de Dominios

Para poder entender las dinámicas evolutivas del modulo efector desde cephalocordados hasta vertebrados, centrados en el grupo intermedio los tunicados, se uso la parsimonia de Dollo para evaluar las perdidas y ganancias. Como se observa en la Figura **3-1**, al evaluar el módulo efector, se observa en las cajas de color blanco la cantidad de dominios necesarios

para llegar a la totalidad de dominios evaluados, es decir que en este caso el ancestro entre cefalocordados y a los Tunicados contaba con 10 como se observa en la caja lo que representa el color lila. El ancestro de los tunicados tuvo una ganancia del dominio de Asticina. La líneas naranjas representan las pérdidas de los dominios en el linaje, este valor influye en el aumento de la caja blanca dentro de los nodos. La caja rellena de color verde representa las ganancias de dominios. Entre otros resultados sobresale que en los tunicados la pérdida de 7 de los 11 dominios evaluados en el módulo efector en la especie Dive, y una pérdida en el clado compuesto por las Cionas y Dive del dominios Distegrin. Por otro lado vemos que tanto para Oidi y Bosc se tienen pérdidasde dominios, tres por parte de Oidi (Colageno, Asticina y Kazal) y en Bosc cuatro dominios (A2M_N, Asticina, Kazal, MAM) como lo muestra la figura **3-1**.

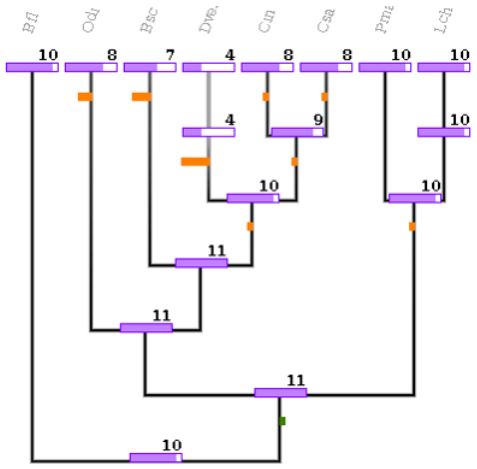

Figura 3-1: Arbol de Dollo evidenciando los cambios de estado en los dominios asociados al Módulo Efector con datos de predicción para *D. vexillum* usando a *C. intestinalis* como referencia y que es mostrada en la figura como Dive-Ciin

Para hacer una comparación de las ganancias y perdidas del módulo efector basados en el modelo de predicción génico de genID con el modelo de Oidi, se usó la parsimonia de Dollo **3-2**. Con estos datos de Dive se observó que se presenta la misma perdida y ganancias de dominios que lo observado para el modelo de predicción de genes basado en Ciin.

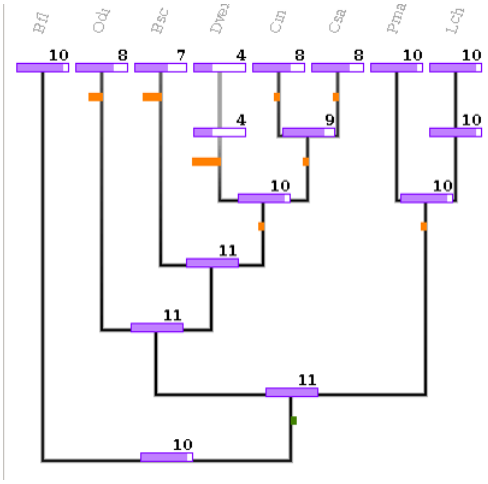

Figura 3-2: Arbol de Dollo evidenciando los cambios de estado en los dominios asociados al Módulo Efector con datos de predicción para *D. vexillum* usando a *O. dioica* como referencia y que es mostrada en la figura como Dive-Oidi

De igual forma se reconstruyeron los modelos de ganancia y perdida en el módulo de señalización en donde se evaluaron 19 dominios. La franja blanca en el ancestro entre los cefalocordados y los tunicados, denota que fue el ancestro de los tunicados que tuvo una ganancia de un dominio en donde se evidencia una pérdida de 11 dominios en Dive, seguida por Bosc con 14, y Oidi con 15. En el grupo de las Cionas se pierden 5 dominios pero Ciin se recupera 7tm_1. Se observa que el dominio NIDO es exclusivo de Vertebrados y que el dominio NACH está ausente en las Cionas, Pema y Oidi. Por otro lado el dominio WH1 y VWA esta ausente tanto en Bosc, Oidi y Dive. Entre las pérdidas más importantes se tiene que W40 está ausente en Bosc, como se observa en la figura **3-3**. Las diferencias observadas en comparación con las predicciones de genes para el modelo de Oidi es que este último modelo no identifica los dominios TSP_1 y WD40 como se observa en la figura **3-4**.

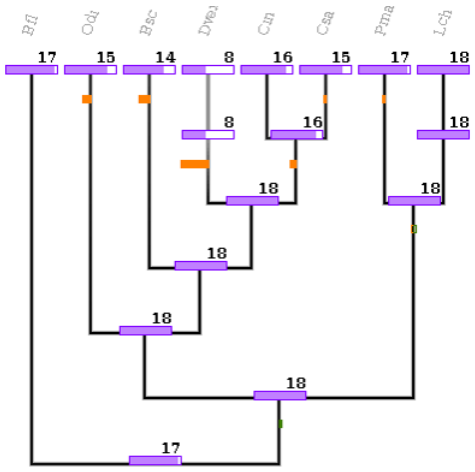

Figura 3-3: Arbol de Dollo para evidenciando los cambios de estado en los dominios asociados al módulo Señalización con datos de predicción para *D. vexillum* usando a *C. intestinalis* como referencia y que es mostrada en la figura como Dive-Ciin

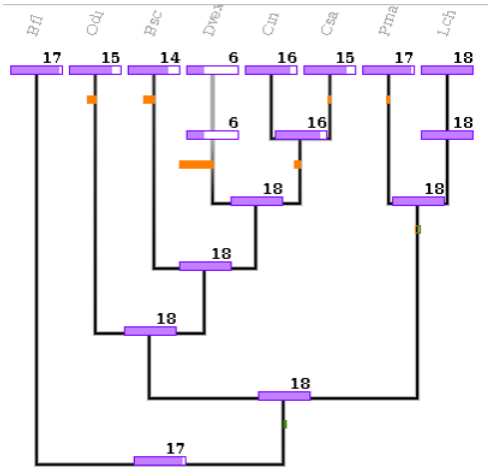

Figura 3-4: Arbol de Dollo para evidenciando los cambios de estado en los dominios asociados al módulo Señalización con datos de predicción para *D. vexillum* usando a *O. dioica* como referencia y que es mostrada en la figura como Dive-Oidi

Por último se evaluó el módulo de reconocimiento, para el cual se encontró que de los 11 dominios evaluados Lach los posee todos, mientras que Dive nuevamente pierde 7 dominios y Oidi 4 dominios. Entre los casos mas interesantes está la perdida del dominio BIR en las Cionas y en Pema, las cadherinas en Dive y Oidi, como se observa en figura **3-5**. En este caso no se observan diferencias en comparación con los dominios calculados para *D. vexillum* usando a *O. dioica* como se ve en la figura **3-6**.

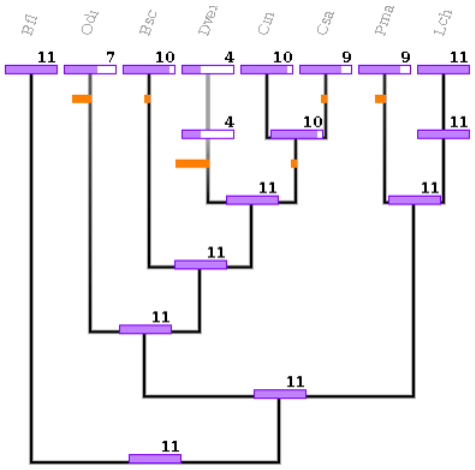

Figura 3-5: Arbol de Dollo para evidenciando los cambios de estado en los dominios asociados al Módulo Reconocimiento con datos de predicción para *D. vexillum* usando a *C. intestinalis* como referencia y que es mostrada en la figura como Dive-Ciin

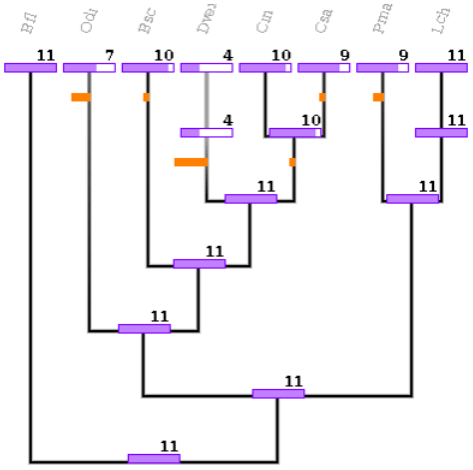

Figura 3-6: Arbol de Dollo para evidenciando los cambios de estado en los dominios asociados al Módulo Reconocimiento con datos de predicción para *D. vexillum* usando a *O. dioica* como referencia y que es mostrada en la figura como Dive-Oidi

Debido a que los tres módulos no logran clasificar todos los dominios asociados al sistema inmune, se evaluaron todos los dominios encontrados por la estrategia ODB independientemente para buscar otro tipo de relaciones como la existencia de dominios específicos por hábitos de vida colonial es decir conocer si están o no asociados algunos módulos del sistema inmune con organismos coloniales o por ejemplo conocer cuántos de éstos están asociados con organismos de vida solitaria y cuáles son los fundamentales para mantenerse en el genoma de

Oidi luego de la reducción genómica. Entonces se generó un diagrama de Venn en donde se muestran todos los dominios compartidos entre los tunicados, usando los dominios obtenidos de los genes putativos resultantes de la estrategia ODB. Ahora, basados en los modelos de arquitectura de genes de Dive predichos con Ciona como referencia, encontramos que hay 48 dominios comunes a todos los Tunicados, dos dominios (PAS_11 y EphA2) compartidos por las especies coloniales. Oidi comparte 47 dominios con las especies solitarias mientras que con las coloniales 19, siendo 19 exclusivos de Oidi, vea la Figura **3-7**, ahora mientras que al ser evaluado los modelos de gen basados en Oidi como referencia, hay una reducción leve en el número de dominios, como por ejemplo el número de dominios compartidos de Dive con los coloniales pasa de 19 a 17 de igual forma con los solitarios, pasa de 47 a 41 y el número total de genes compartidos por todos los tunicados desciende de 48 a 42 como se observa en la Figura **3-8**.

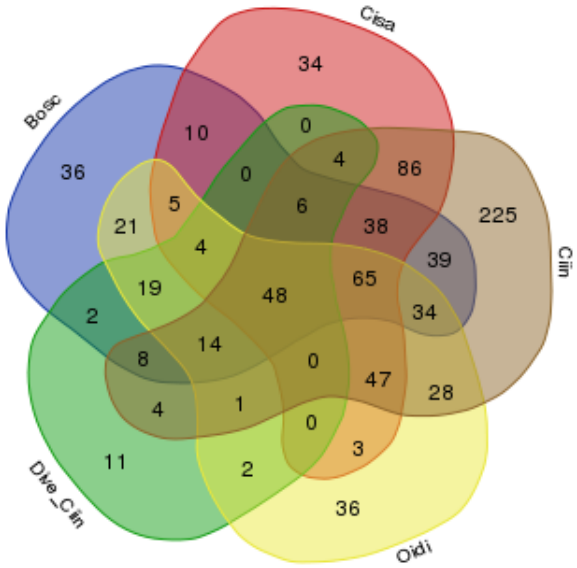

Figura 3-7: Diagrama de Venn donde se compara los dominios compartidos entre Dive _Ciin y el resto de tunicados

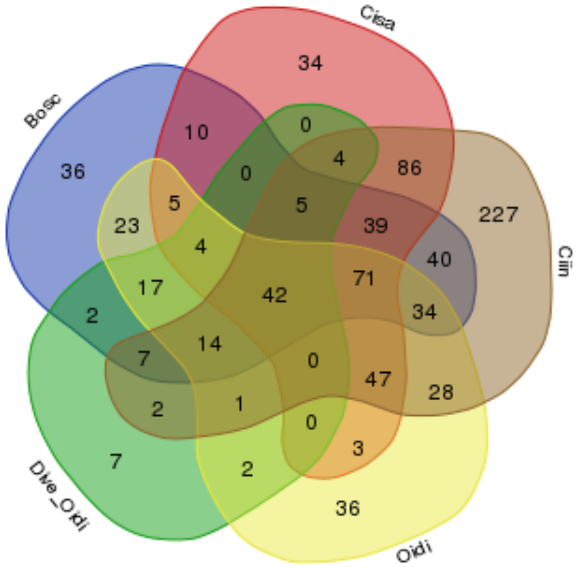

Figura 3-8: Diagrama de Venn donde se compara los dominios compartidos entre Dive_Oidi y el resto de tunicados

Una vez establecidos los dominios de tunicados se evaluó por medio de otro diagrama de Venn el comportamiento de los dominios en el clado de los vertebrados, con miras a poder establecer algunos dominios exclusivos de tunicados o algunas perdidas en los grupos externos, tanto vertebrados como cefalocordados. Para ello tomamos los 48 dominios resultantes de la gráfica **3-7** compartidos entre todos los tunicados y el grupo externo a túnicados (ver gráfica **3-9**). Se observa que de los 48, 42 son compartidos por todos los grupos externos, mientras que solo 6 son compartidos por Cefalocordados, tunicados y Lach, con una ausencia evidente de algunos de estos dominios conservados en Tunicados que se encuentran perdidos en Vertebrados.

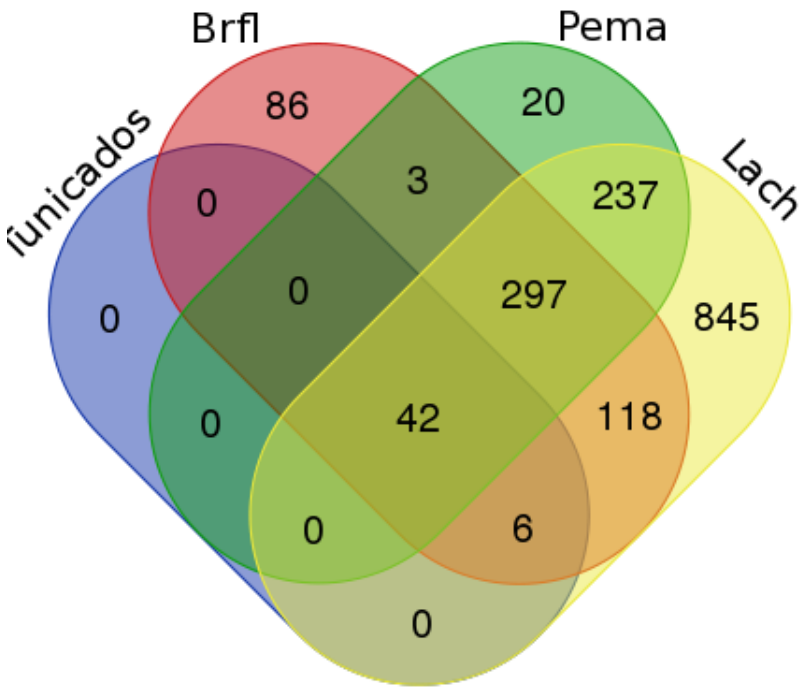

Figura 3-9: Diagrama de Venn donde se compara los dominios compartidos entre Los dominios compartidos entre los tunicados (Oidi, Dive, Bosc, Ciin y Cisa) contra los dominios totales de los Vertebrados y Brfl

Con miras a identificar dominios ancestrales en la colonialidad en Dive, evaluamos los dominios compartidos entre Dive y Brfl contra el resto de tunicados **3-10**, de esos 18 dominios, encontramos 17 están ausentes en las Cionas mientras que tan solo uno es compartido por (Bosc, Dive, Brf, Oidi y Ciin).

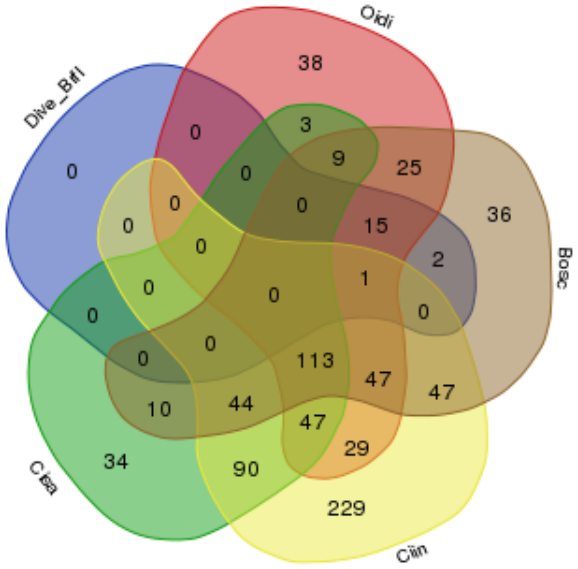

Figura 3-10: Diagrama de Venn donde se compara los dominios compartidos entre Dive y Brfl, contra los dominios totales del resto de tunicados

Por último con miras a esclarecer cual es el comportamiento evolutivo de los dos dominios encontrados como exclusivos de organismos coloniales provenientes de la figura **3-7** (PAS_11 y EphA2), se decidió evaluar solo estos dos dominios con los grupos externos y se encontró que son dominios ancestrales compartidos con Brfl pero ausentes en vertebrados. ver figura **3-11**

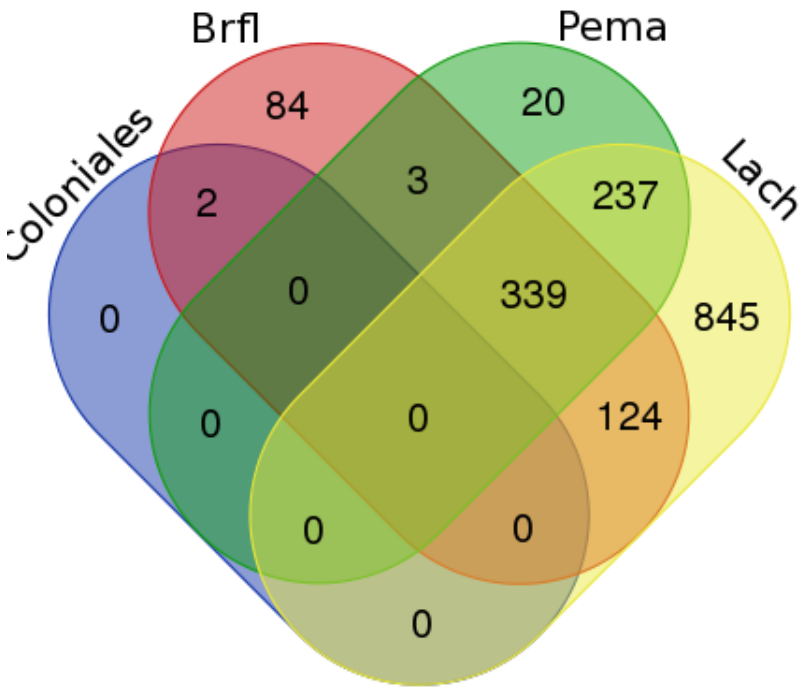

Figura 3-11: Diagrama de Venn donde se compara los dominios compartidos los Tundeados Coloniales Dive y Boch, contra los dominios totales del Vertebrados y Brfl

3.3.2. Relaciones de ortologia entre las proteínas de la estrategia ODB

Con el fin de establecer las relaciones de Ortología, se usó ProteinOrtho según los parámetros descritos en la metodología. Los resultados de relaciones de ortología entre las especies se muestra la gráfica **3-12**. En este diagrama las flechas representan relaciones de ortología establecidas entre las proteínas derivadas de la estrategia ODB cuantificadas desde las salidas de las relaciones de ortología detectadas por ProteinOrtho para cada par de especies estudiadas. En la parte superior de cada una de las líneas se reporta el número de relaciones de ortología 1-1 mientras que en la línea de abajo se reporta el número de candidatos coortologos. Se destaca la amplia cantidad de relaciones de ortología entre las especies Lach y Brfl correspondientes a 25 relaciones de ortología hasta 28 relaciones con Pema y 16 coortólogos entre Pema y Brfl mientras que Bosc se tiene como máximo 11 relaciones de ortologia.

Por otro lado la especie Oidi presenta un máximo de relaciones de ortología con Lach de 15 y 5 secuencias de ortólogos y Dive es el que menos relaciones de ortología se encontraron siendo con Oidi y Ciin el máximo con 2, y por último se encontró que Ciin y Cisa tienen el mayor número de relaciones de ortologia entre ellos con 18 para Cisa con Ciin, mientras que Ciin presenta 26 con Lach.

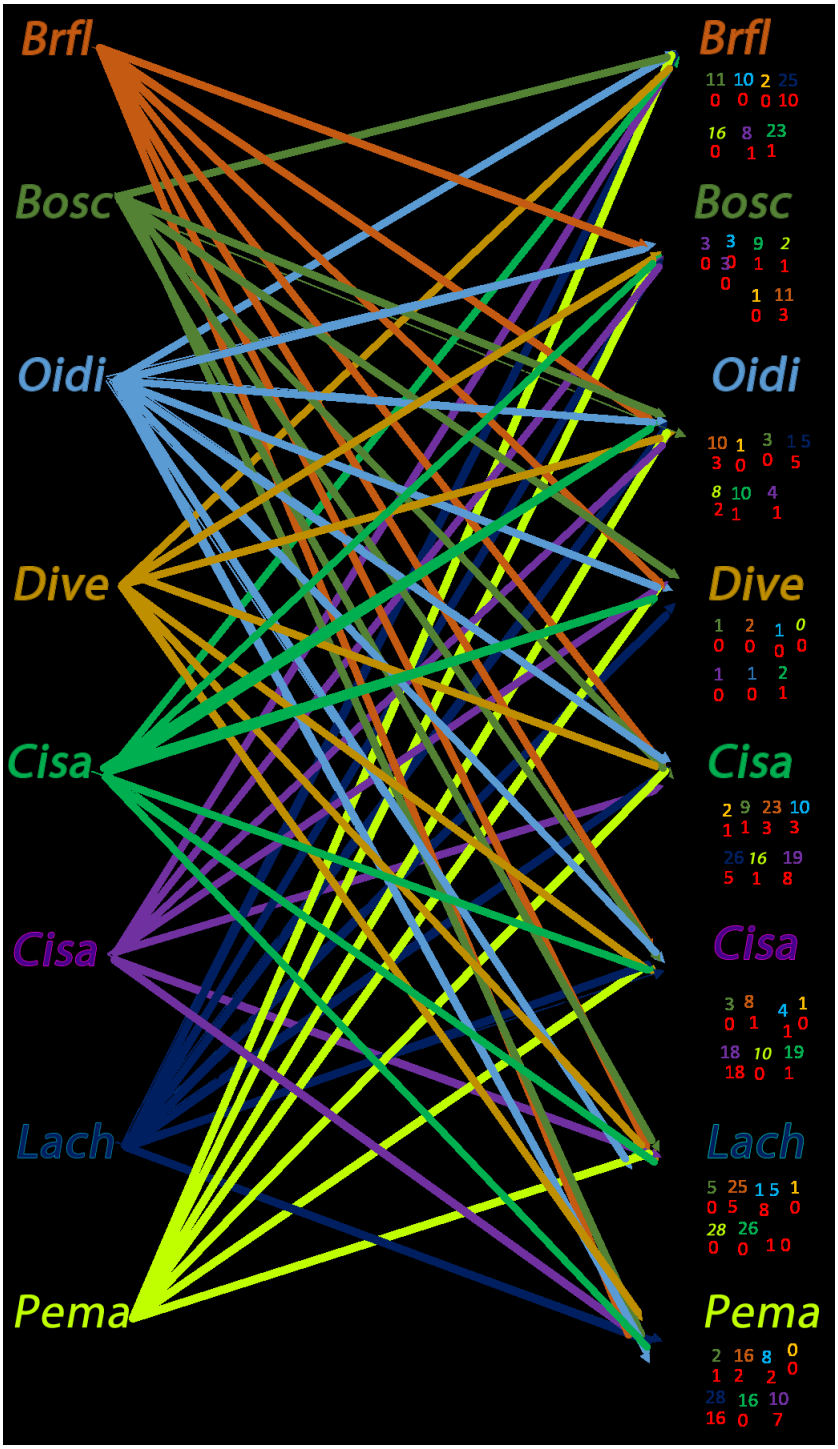

Figura 3-12: Este diagrama muestra las relaciones de ortología (número en la parte superior de la flecha) y de coortologia (número inferior en la flecha) entre las diferentes proteínas anotadas por ODB

3.4. Discusión

Al momento de evaluar los dominios según los dos modelos de genes propuestos por geneID para Dive, uno basado en un modelo que usa de referencia a Ciin y el otro basado en referencia de Oidi, podemos concluir que estos dos modelos son complementarios y que debido a la cercanía en la estructura del genoma de Ciin con Dive (menos compacto y con más información que Oidi) era de esperarse que tuviese mayor número de predicciones, las cuales pudieron haber estado ausentes debido a la predicción del la estructura del gen basado en modelos de genes por referencia de Oidi. Sin embargo aunque existen diferencias por cada uno de los módulos estas no son extremas tanto en la cantidad de genes como en la cantidad de dominios detectados.

Debido a la poca información existente sobre la naturaleza de los dominios asociados al sistema inmune en tunicados, nos centramos de una forma conservadora en la estrategia ODB para recuperar el aporte de todas las especies del *Gold Standard* al soporte de las ganancias y perdidas de dominios. Entonces el modelo de pérdida y ganancia de Dollo fue aplicado sobre los dominios más conservados y establecidos con evidencia en literatura su asociación a estos módulos propuesta por [122].

Al observar las pérdidas y ganancias de dominios usando la parsimonia de Dollo, observamos que tanto en el módulo efector como en el módulo señalización, se dió la ganancia de un nuevo dominio en el ancestro de los tunicados que se mantuvo hasta los vertebrados. También se logra ver una mayor pérdida recambio de dominios o en el módulo efector en los tunicados pero más conservado en vertebrados. Este fenómeno se puede deber a que las rutas efectoras se estabilizaron en los vertebrado después del big ban inmunológico, mientras que al tener diferentes hábitat y modos de vida, los tunicados variaron sus rutas acomodándolas a sus necesidades inmunológicas. Por otro lado, en el módulo de señalización, la parsimonia de Dollo nos muestra que aunque hubo pérdidas no fueron tan notorias como en el módulo efector y que tanto en vertebrados como en tunicados es mucho más estable, denotando que la señalización del sistema inmune es mucho más ancestral y estable, lo cual no da mucho espacio de maniobra para que en la evolución se hubiesen explorado alternativas inmunológicas.

Caso similar ocurre con el módulo de reconocimiento, donde se denota un comportamiento más ancestral ya que todos los dominios son rastreables hasta Brfl, aunque cabe resaltar la pérdida de cuatro (Fbox, Cadherina, integrinas y SRCR) dominios en Oidi aunque este fenómeno de perdida de dominios es comparable con los dos módulos previamente mencionados (señalización y efector). Se observa un máximo de pérdida de dos dominios del set original proveniente de los cefalocordados.

Encontramos que en todos los módulos existe el patrón de pérdidas sobresalientes de dominios en la especie *Didemnum vexillum*, lo cual puede deberse a la calidad del ensamblaje genómico de la especie que conlleva a una predicción de genes y por ende de proteínas no óptima. Sin embargo, el modelo cuando se analiza en contraste con el resto de organismos, nos señala

que el funciona ya que logra predecir dominios de proteínas putativas asociadas al sistema inmune y que en el rastreo de dominios no se presentaron cambios radicales al evaluarse la ganancia incluyendo a los dominios de Dive detectados con una u otra referencia.

De igual forma, como se ve en el análisis de la parsimonia de Dollo, los dominios ancestrales provenientes de los cefalocordados son compartidos con tunicados, e incluso con los vertebrados. El caso del dominio Asticina es interesante ya que fue una ganancia para el sistema inmune de tunicados y el cual se ha mantenido hasta los vertebrados. Este tipo de hallazgos nos puede sugerir que no solo el sistema inmune está basado en dominios altamente conservados sino que puede integrar nuevos dominios para que trabajen con los dominios constitutivos y de esta forma explorar rutas alternativas de inmunidad que respondan a las necesidades de cada organismo, generando de esta forma casi que un sistema inmune único y especifico a cada especie.

Con nuestra metodología es posible generar candidatos de proteínas putativas del sistema inmune basados en una estrategia que mantiene la flexibilidad de búsqueda de dominios. Ahora, ya que los tunicados son un grupo importante no solo para el entendimiento del surgimiento de la inmunidad adaptativa y sino para comprender procesos claves en la biología como aleoreconocimiento se observa según los análisis anteriores que la modularidad persiste y que se encuentran dominios claves de la inmunidad según los análisis de intersección de dominios presentados en los diagrams de Venn. En este caso son 48 para ser exactos y podemos observar que estos se mantienen desde los cefalocordados y que mantiene su función ya sea en tunicados solitarios como las Cionas, o en organismos con pérdidas masivas de información genética como Oidi o en el otro extremo organismos de vida colonial, cada una deseas especies presentan por su puesto necesidades inmunológicas diferentes que aun así mantienen estos dominios y son rastreables hasta los vertebrados.

Uno de los resultados más interesantes e intrigantes producto de esta comparación básica, es la separación de los sistema inmunes de las Cionas con respecto al resto de Tunicados, por ejemplo una especie con pérdidas masivas de información genética como Oidi comparte más dominios con los organismos coloniales que con los solitarios. Este fenómeno se suma a la gran cantidad de dominios únicos en la especies Ciin observados en la gráfica **3-7** donde se encuentran 225 Dominios no compartidos entre Ciin y los otros tunicados, y a la perdida de dominios de gran importancia en los diferentes módulos como el de reconocimiento (BIR) y el de señalización (NACH).

Otro resultado que merece la atención, son los dominios PAS_11 y EphA2 los cuales se pueden encontrar en Brfl y en tunicados coloniales pero ausente en vertebrados y en los tunicados solitarios, sugiriendo este fenómeno que estos genes pueden ser excelentes candidatos a tener un rol importante en el alorreconocimiento, y quizás en poder rastrear algún tipo de sintenia entre esta función tan importante.

Es importante resaltar que antes de iniciar este trabajo se tenía claro que existían 9 dominios constitutivos al sistema inmune según [88], pero nosotros logramos encontrar 48 dominios que se encuentran a lo largo de todas las especies evaluadas. Este fenómeno era de esperarse

ya que la combinatoria de dominios y el exonshufling que pueden ser motor fundamental de la variabilidad de los diferentes sistemas inmunes podrían explicar nuestros resultados. Adicionalmente el trabajo reportado por [88] en el 2007 no incorporaba muchas de las especies de tunicados presentes en este trabajo ni la definición de modularidad descrita para los tunicados según [122].

A pesar de tener un repertorio establecido de dominios cada especie, los fenómenos de exónshufling, la incorporación de nuevos dominios y la plasticidad que solo un sistema modular como lo es el sistema inmune permite generar un universo inmunológico capaz de responder a las necesidades adaptativas de cada organismo. Por último partimos de la hipótesis que Oidi era un organismo con pérdidas genómicas importantes, por lo que se esperaba que el repertorio de dominios fuera muy limitado, pero observamos que el repertorio no se redujo sino que se sacrificó quizás en el número de copias y la diversidad de arquitecturas. Lo que sugiere que Oidi ha respondido a sus necesidades inmunológicas con receptores con baja complejidad en cuanto a su arquitectura pero con una gran diversidad de dominios.

Por otro lado, del estudio de Ortologia realizado por ProteinOrtho. Es importante aclarar que este método es basado en una estrategia de Blast recíproco y principios de teoría de grafos. Ya que el método es basado en Blast se tienen limitaciones para la detección de ortólogos que no compartan alto grado de similitud entre secuencias de proteínas completas. Es por ello que se seleccionaron proteínas para este estudio las proteínas que soportan la detección de dominios por la estrategia ODB, desafortunadamente aun carecemos de un método directo para la predicción de ortólogos diferente a un método basado en secuencia completa de la proteína. Teniendo en cuenta esta restricción, no todas las relaciones de ortología fueron detectables. Con este método se demostró que las proteínas con mayor número de ortólogos fueron las provenientes de los grupos externos Brfl y Lach, y dado la restricción de capturar ortologías basadas directas en secuencia de proteínas no todas relaciones de ortología fueron detectadas en los tunicados. Se observa que las relaciones de ortología de proteínas de grupo ODB dentro del grupo de tunicados no fueron tan fuertes.

Es el caso de Bosc el cual a diferencia de Oidi que no tiene un genoma compacto, el número d observado de relaciones de ortología con el resto de tunicados, al igual que en Dive, es muy limitado. Esto se puede deber a que al presentar un estilo de vida colonial el sistema inmune se ha tornado muy específico dejando un repertorio limitado de proteínas ortólogas y dando paso al nacimiento y conformación de nuevas arquitecturas que permitan afrontar la colonialidad y que pueden estar dentro del grupo de estrategias de Desorden que no pueden ser detectadas en el método de ProteinOrtho.

La presencia de un mayor número de proteínas ortólogas en Oidi, nos corroboró que aunque hay una perdida masiva de información genética en Oidi las estructuras básicas se conservan y su evolución es rastreable hacia los grupos externos como Lach y Pema. Este análisis demostró que a pesar que Lach sea más distante evolutivamente que Pema para los tunicados, la expansión de estructuras genéticas permitieron mejores resultados al momento de rastrear ortólogos siendo un mejor referente que esta especie filogenéticamente más cercana. Sorpren-

de ver el número de copias de cada gen ortólogo que presenta Cisa ya que llega presentar hasta 18 (el número más alto) contra Lach, mostrando la alta similaridad y cercanía que tiene las cionas sobre el sistema inmune de los vertebrados, mucho más que las arquitecturas presentes en otros tunicados. Sin embargo es importante tener presente que estudios anteriores sugieren que en los tunicados las tasas de evolución son altas y por tanto se esperan patrones de organización específicos de todos sus genes[94, 10]. Eventos particulares en la evolución de genomas de estas especies pueden verse reflejados en sistemas complejos como el sistema inmune como se observa en estas proteínas ultra conservadas.

4 Conclusiones

- Las metodologías propuestas permitieron cumplir con cada uno de los objetivos, es decir primero delimitar los dominios componentes del sistema inmune utilizando estrategias alternativas a la aproximación básica de comparación de similitud de secuencias completas de genes o de proteínas o a las aproximaciones ampliamente usadas como GO terms. Así la estrategia de definir un sistema de referencia de dominios *Gold Standard* es una novedosa estrategía que permite rastrear genes en otras especies basados en su arquitectura. Asi fue posible establecer un conjunto real de Dominios asociados al sistema inmune que permitan explorar arquitecturas variables y fleixbles esperadas en la evolución de los componentes del sistema inmune de los tunicados.

- Se estableció una metodología para rastrear fenómenos de ganancia y perdida de dominios, logrando demostrar que la evolución del sistema inmune es un problema real al momento de intentar anotar genes del sistema inmune en especies distantes evolutivamente de los organismos modelo para estudiar el sistema inmune como son los mamíferos. Este tipo de aproximación permitió detectar dominios específicos y compartidos entre los grupos de estudio que no son rastreables por las estrategias ampliamente usadas como la homología directa. Por tanto el uso de dominios recuperados desde la arquitectura *Gold standard* nos permitió enfocarnos en los modelos funcionales de la arquitectura de la proteina mas allá del uso clásico de la similitud de secuencias. El uso de la estrategia de Desorden le da flexibilidad el modelo ya que captura en la estructura fenómenos que puedan estar sometidas a eventos de exón shufflin, logrando llegar a anotar proteínas que por homología directa no se podrían anotar.

- Este estudio ha permitido entender fenómenos de la evolución de dominios del sistema inmune en la esencia del mismo, y es la naturaleza modular del sistema inmune que persiste en los grupos de estudio. Logramos a través de esta metodología confirmar que el sistema inmune es un sistema modular dinámico, el cual está basado en la estructura y combinación de dominios y forma una base para la evolución de respuestas inmunes alternantes que pueden estar alimentadas por la ganancia o perdida dichas de dominios. Sin embargo se resalta que los resultados indican que existe un conjunto base o mínimo de dominios que permiten moldear una respuesta casi que especifica a las necesidades inmunológicas de cada especie de tunicado.

- Durante este trabajo se logró establecer que a partir las dinámicas evolutivas de los

dominios, éstas están ampliamente relacionadas con el nivel de compactación del genoma, lo cual se observó en la especie *O. dioica* (Oidi), sin embargo este fenómeno de evolución genómica no limitó su repertorio de proteínas asociadas al sistema inmune pero si en la cantidad de repeticiones de cada dominio generando una arquitectura en las proteínas específicas y otras compartidas con otras especies de tunicados.

- Por último, podemos decir que nuestra metodología logra proponer una posible solución básica al problema del aloreconocimieto ya que logramos identificar dominios asociados al sistema inmune que son exclusivos de las especies coloniales, lo cual da una luz en el arduo camino para entender los componentes y las dinámicas del alorreconocimiento en tunicados.

- Se demostró que para especies que no son modelo, la mejor forma de abordar la evolución del sistema inmune es por medio del estudio de los dominios de la componen, y con diferentes aproximaciones que permitan flexibilidad el momento de anotar sus arquitecturas proteicas.

5 Anexos

5.1. Capítulo 1: Construcción de Pipeline automatizada para la identificación de dominios asociados a proteínas del sistema inmune de Tunicados

En el primer conjunto se observa cuantas estructuras canónicas de *Acyrthosiphon pisum* pudieron ser identificadas en los diferentes tunicados, se observa que hay una tendencia en las estructuras canónicas de especies provenientes de Ensembl a tener un mejor desempeño por medio de la estrategia **D**esorden y la combinatoria entre **O**rden/**D**rden (**OD**), acompañada de un buen número de predicciones por medio de la estrategia **O**rden en las especies cuyos genomas no provienen de ensembl *Brfl*,*Bosc* y textsf$Oidi$, estas mismas tendencias se observan al evaluar las arquitecturas canónicas de *Mus musculus* y *Homo sapiens*, aunque en el cuadro donde se evalúa la arquitectura de *Acpi* se observa una diferencia significativa en comparación con los mamíferos, y es un mayor número de anotaciones por medio de la estrategia **ODB**

En contraposición se observa que las arquitecturas canónicas de las especies *Nasonia vitripennis*, *Apis mellifera* y *Anopheles gambiae*, presentan tendencias similares, donde se destaca la presencia de las estrategias **ODB** en alta cantidad, siendo más preponderante las frecuencias de esta estrategia en *Lach*, y a su vez se pueden observar que tanto *Ciin*, *Cisa*, y *Pema* tienen una anotación considerable por la misma estrategia. La combinatoria de las estrategias de **B**last y **O**rden, mostraron ser útiles en los genomas que no proviene de Ensembl, aunque en menor medida para textsf$Apis\ mellifera$, en contraposición a lo observado en las especies *Acyrthosiphon pisum*, *Mus musculus* y *Homo sapiens*, hay presencia de anotaciones por medio de la estrategia **DB**, pero una reducción considerable en la estrategia de **O**rden para las especies *Anopheles gambiae* y *Apis mellifera* pero se ve compensada por la estrategia **OB** ya mencionada.

Por último se observa que las estructuras canónicas procedentes de *Drosophila melanogaster* tienen una tendencia intermedia en comparaciones las dos tendencias previamente escritas, ya que hay presencia tanto de **ODB**, **D** y de **OD**, esta ultima ausente en los analisis de *Acyrthosiphon pisum*, *Mus musculus* y *Homo sapiens* pero tiene similaridad con este grupo en cuanto a que hay ausencia de predicciones por medio de las estrategias **OB** y **O**.

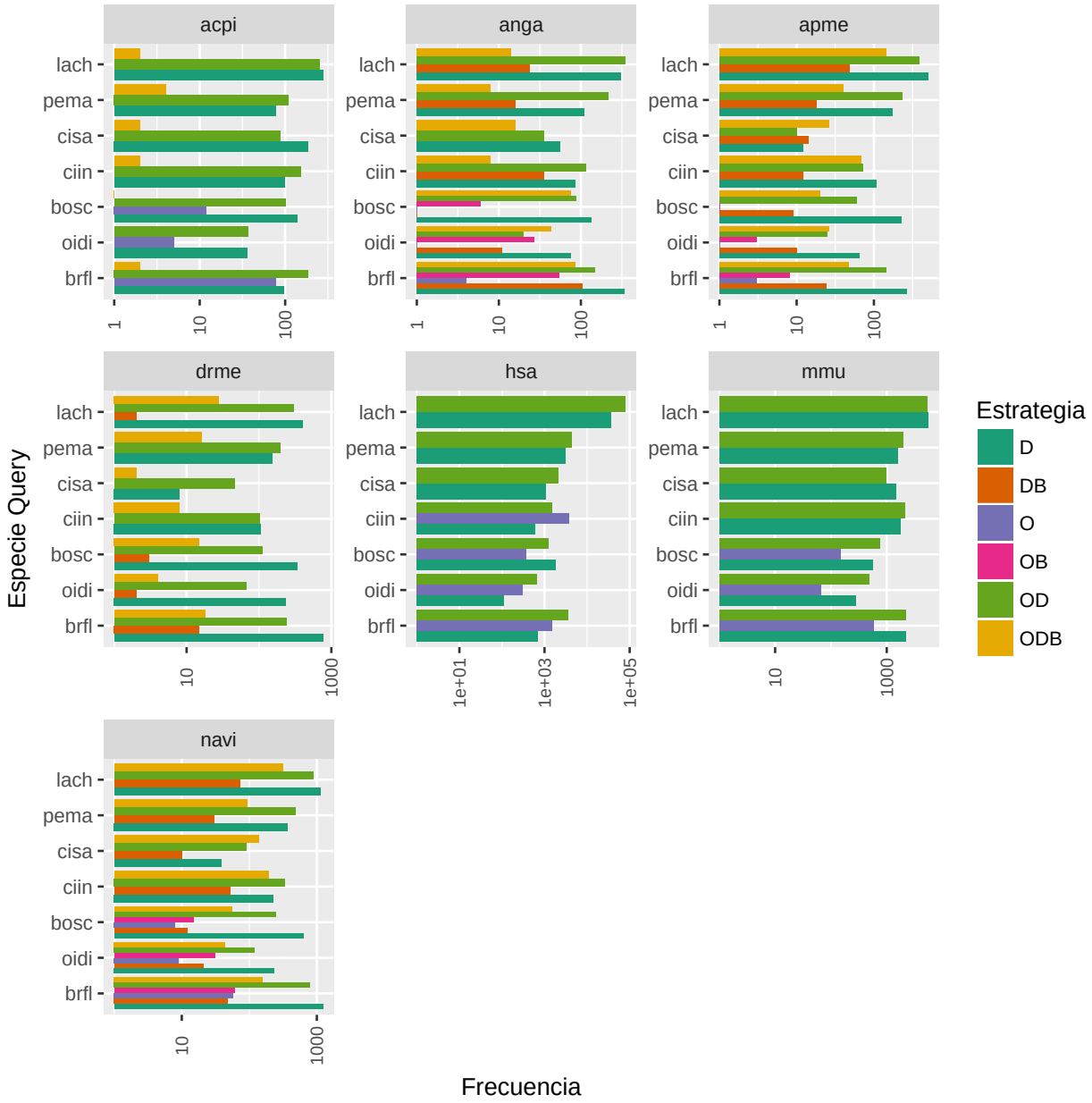

Figura 5-1: Frecuencias de las diferentes estrategias de anotación: **ODB**, en las especies de referencia *Mus musculus*, *Homo sapiens*, *Nasonia vitripennis*, *Apis mellifera*, *Drosophila melanogaster*, *Anopheles gambiae* y *Acyrthosiphon pisum*. con las que se anotaron los genes del SII en cada una los cordados objeto de estudio

Bibliografía

[1] Aken, Bronwen L. ; Ayling, Sarah ; Barrell, Daniel ; Clarke, Laura ; Curwen, Valery ; Fairley, Susan ; Banet, Julio F. ; Billis, Konstantinos ; Girón, Carlos G. ; Hourlier, Thibaut [u. a.]: The Ensembl gene annotation system. En: *Database* 2016 (2016), p. baw093

[2] Albalat, R. ; Canestro, C.: Evolution by gene loss. En: *Nat. Rev. Genet.* 17 (2016), 07, Nr. 7, p. 379–391

[3] Alberts, Bruce: *Molecular biology of the cell*. Garland science, 2017

[4] Altincicek, B. *REVIEW The innate immunity in the cnidarian Hydra vulgaris*. 2009

[5] Antón Marín, Yanet: El sistema inmune de los invertebrados-The immune.

[6] Ashburner, Michael ; Ball, Catherine A. ; Blake, Judith A. ; Botstein, David ; Butler, Heather ; Cherry, J M. ; Davis, Allan P. ; Dolinski, Kara ; Dwight, Selina S. ; Eppig, Janan T. [u. a.]: Gene Ontology: tool for the unification of biology. En: *Nature genetics* 25 (2000), Nr. 1, p. 25–29

[7] Bailey, Mick ; Christoforidou, Zoe ; Lewis, Marie: Evolution of immune systems: Specificity and autoreactivity. En: *Autoimmunity Reviews* 12 (2013), Nr. 6, p. 643 – 647. – Natural Antibodies in Health and Disease. – ISSN 1568–9972

[8] Barrett, Lucy W. ; Fletcher, Sue ; Wilton, Steve D.: Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. En: *Cellular and molecular life sciences* 69 (2012), Nr. 21, p. 3613–3634

[9] Bayne, Christopher J.: Origins and Evolutionary Relationships Between the Innate and Adaptive Arms of Immune Systems. En: *Integrative and Comparative Biology* 43 (2003), Apr, Nr. 2, p. 293–299. – ISSN 1540–7063

[10] Berná, Luisa ; Alvarez-Valin, Fernando: Evolutionary genomics of fast evolving tunicates. En: *Genome biology and evolution* 6 (2014), Nr. 7, p. 1724–1738

[11] Bhattacharya, Sanchita ; Andorf, Sandra ; Gomes, Linda ; Dunn, Patrick ; Schaefer, Henry ; Pontius, Joan ; Berger, Patty ; Desborough, Vince ; Smith, Tom ; Campbell, John [u. a.]: ImmPort: disseminating data to the public for the future of immunology. En: *Immunologic research* 58 (2014), Nr. 2-3, p. 234–239

[12] Bianchi, Marco E.: DAMPs, PAMPs and alarmins: all we need to know about danger. En: *Journal of Leukocyte Biology* 81 (2007), Nr. 1, p. 1–5

[13] Birney, Ewan ; Andrews, T D. ; Bevan, Paul ; Caccamo, Mario ; Chen, Yuan ; Clarke, Laura ; Coates, Guy ; Cuff, James ; Curwen, Val ; Cutts, Tim [u. a.]: An overview of Ensembl. En: *Genome research* 14 (2004), Nr. 5, p. 925–928

[14] Birney, Ewan ; Clamp, Michele ; Durbin, Richard: GeneWise and genomewise. En: *Genome research* 14 (2004), Nr. 5, p. 988–995

[15] Blanco, Enrique ; Parra, Genís ; Guigó, Roderic: Using geneid to identify genes. En: *Current protocols in bioinformatics* (2007), p. 4–3

[16] Boehm, Thomas: Design principles of adaptive immune systems. En: *Nature Reviews Immunology* 11 (2011), Nr. 5, p. 307–317

[17] Boehm, Thomas ; Iwanami, Norimasa ; Hess, Isabell: Evolution of the immune system in the lower vertebrates. En: *Annual review of genomics and human genetics* 13 (2012), p. 127–149

[18] Breuer, Karin ; Foroushani, Amir K. ; Laird, Matthew R. ; Chen, Carol ; Sribnaia, Anastasia ; Lo, Raymond ; Winsor, Geoffrey L. ; Hancock, Robert E. W. ; Brinkman, Fiona S. L. ; Lynn, David J.: InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. En: *Nucleic Acids Research* 41 (2013), Nr. D1, p. D1228–D1233

[19] Breuer, Karin ; Foroushani, Amir K. ; Laird, Matthew R. ; Chen, Carol ; Sribnaia, Anastasia ; Lo, Raymond ; Winsor, Geoffrey L. ; Hancock, Robert E. ; Brinkman, Fiona S. ; Lynn, David J.: InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. En: *Nucleic acids research* (2012), p. gks1147

[20] Brucker, Robert M. ; Funkhouser, Lisa J. ; Setia, Shefali ; Pauly, Rini ; Bordenstein, Seth R.: Insect Innate Immunity Database (IIID): An Annotation Tool for Identifying Immune Genes in Insect Genomes. En: *PLOS ONE* 7 (2012), 09, Nr. 9, p. 1–4

[21] Brusca, RC ; Brusca, GJ: *Invertebrates 2nd Eds.* Sinauer, Associates, Sunderland, Mass, 2003

[22] Buckley, Katherine M. ; Rast, Jonathan P.: Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. En: *Developmental & Comparative Immunology* 49 (2015), Nr. 1, p. 179 – 189. – ISSN 0145–305X

[23] Burge, C. ; Karlin, S.: Prediction of complete gene structures in human genomic DNA. En: *J. Mol. Biol.* 268 (1997), Apr, Nr. 1, p. 78–94

[24] Burge, Chris ; Karlin, Samuel: Prediction of complete gene structures in human genomic DNA. En: *Journal of molecular biology* 268 (1997), Nr. 1, p. 78–94

[25] Calvano, Steve E. ; Xiao, Wenzhong ; Richards, Daniel R. ; Felciano, Ramon M. ; Baker, Henry V. ; Cho, Raymond J. ; Chen, Richard O. ; Brownstein, Bernard H. ; Cobb, J P. ; Tschoeke, S K. [u. a.]: A network-based analysis of systemic inflammation in humans. En: *Nature* 437 (2005), Nr. 7061, p. 1032–1037

[26] Cambi, Alessandra ; Koopman, Marjolein ; Figdor, Carl G.: How C-type lectins detect pathogens. En: *Cellular Microbiology* 7 (2005), Nr. 4, p. 481–488. – ISSN 1462–5822

[27] Cantarel, B. L. ; Korf, I. ; Robb, S. M. ; Parra, G. ; Ross, E. ; Moore, B. ; Holt, C. ; Sanchez Alvarado, A. ; Yandell, M.: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. En: *Genome Res.* 18 (2008), Jan, Nr. 1, p. 188–196

[28] Carroll, Sean B.: Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. En: *Cell* 134 (2008), Nr. 1, p. 25–36

[29] Claverie, Jean-Michel: Computational Methods for the Identification of Genes in Vertebrate Genomic Sequences. En: *Human Molecular Genetics* 6 (1997), Nr. 10, p. 1735

[30] Cooper, Max D. ; Alder, Matthew N.: The Evolution of Adaptive Immune Systems. En: *Cell* 124 (2006), Nr. 4, p. 815 – 822. – ISSN 0092–8674

[31] Csűös, Miklós: Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. En: *Bioinformatics* 26 (2010), Nr. 15, p. 1910–1912

[32] Dam, Tarun K. ; Brewer, C F.: Lectins as pattern recognition molecules: The effects of epitope density in innate immunity*. En: *Glycobiology* 20 (2010), Nr. 3, p. 270–279

[33] Dehal, Paramvir ; Satou, Yutaka ; Campbell, Robert K. ; Chapman, Jarrod ; Degnan, Bernard ; De Tomaso, Anthony ; Davidson, Brad ; Di Gregorio, Anna ; Gelpke, Maarten ; Goodstein, David M. [u. a.]: The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. En: *Science* 298 (2002), Nr. 5601, p. 2157–2167

[34] Denoeud, France ; Henriet, Simon ; Mungpakdee, Sutada ; Aury, Jean-Marc ; Da Silva, Corinne ; Brinkmann, Henner ; Mikhaleva, Jana ; Olsen, Lisbeth C. ;

Jubin, Claire ; Cañestro, Cristian [u. a.]: Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. En: *Science* 330 (2010), Nr. 6009, p. 1381–1385

[35] Denoeud, France ; Henriet, Simon ; Mungpakdee, Sutada ; Aury, Jean-Marc ; Da Silva, Corinne ; Brinkmann, Henner ; Mikhaleva, Jana ; Olsen, Lisbeth C. ; Jubin, Claire ; Cañestro, Cristian [u. a.]: Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. En: *Science* 330 (2010), Nr. 6009, p. 1381–1385

[36] Dufayard, J. F. ; Duret, L. ; Penel, S. ; Gouy, M. ; Rechenmann, F. ; Perriere, G.: Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. En: *Bioinformatics* 21 (2005), Jun, Nr. 11, p. 2596–2603

[37] Durinck, Steffen ; Moreau, Yves ; Kasprzyk, Arek ; Davis, Sean ; De Moor, Bart ; Brazma, Alvis ; Huber, Wolfgang: BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. En: *Bioinformatics* 21 (2005), Nr. 16, p. 3439–3440

[38] Durinck, Steffen ; Spellman, Paul T. ; Birney, Ewan ; Huber, Wolfgang: Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. En: *Nat. Protocols* 4 (2009), 07, Nr. 8, p. 1184–1191. ISBN 1754–2189

[39] Eddy, Sean R.: Profile hidden Markov models. En: *Bioinformatics* 14 (1998), Nr. 9, p. 755–763

[40] Famiglietti, Maria L. ; Estreicher, Anne ; Gos, Arnaud ; Bolleman, Jerven ; Géhant, Sébastien ; Breuza, Lionel ; Bridge, Alan ; Poux, Sylvain ; Redaschi, Nicole ; Bougueleret, Lydie [u. a.]: Genetic Variations and Diseases in UniProtKB/Swiss-Prot: The Ins and Outs of Expert Manual Curation. En: *Human mutation* 35 (2014), Nr. 8, p. 927–935

[41] Fang, G. ; Bhardwaj, N. ; Robilotto, R. ; Gerstein, M. B.: Getting started in gene orthology and functional analysis. En: *PLoS Comput. Biol.* 6 (2010), Mar, Nr. 3, p. e1000703

[42] Felsenstein, Joseph ; Felenstein, Joseph: *Inferring phylogenies*. Vol. 2. Sinauer associates Sunderland, MA, 2004

[43] Finn, Robert D. ; Clements, Jody ; Eddy, Sean R.: HMMER web server: interactive sequence similarity searching. En: *Nucleic acids research* (2011), p. gkr367

[44] Finn, Robert D. ; Coggill, Penelope ; Eberhardt, Ruth Y. ; Eddy, Sean R. ; Mistry, Jaina ; Mitchell, Alex L. ; Potter, Simon C. ; Punta, Marco ; Qureshi, Matloob ; Sangrador-Vegas, Amaia ; Salazar, Gustavo A. ; Tate, John ; Bateman, Alex: The Pfam protein families database: towards a more sustainable future. En: *Nucleic Acids Research* 44 (2016), Nr. D1, p. D279–D285

[45] Flajnik, M. F. ; Kasahara, M.: Origin and evolution of the adaptive immune system: genetic events and selective pressures. En: *Nat. Rev. Genet.* 11 (2010), Jan, Nr. 1, p. 47–59

[46] Flajnik, Martin F. ; Kasahara, Masanori: Origin and evolution of the adaptive immune system: genetic events and selective pressures. En: *Nature Reviews Genetics* 11 (2010), Nr. 1, p. 47

[47] Franzenburg, Sören ; Fraune, Sebastian ; Künzel, Sven ; Baines, John F. ; Domazet-Lošo, Tomislav ; Bosch, Thomas C. G.: MyD88-deficient Hydra reveal an ancient function of TLR signaling in sensing bacterial colonizers. En: *Proceedings of the National Academy of Sciences* 109 (2012), Nr. 47, p. 19374–19379

[48] Gilbert, Scott F.: Genes classical and genes developmental. En: *The concept of the gene in development and evolution* (2000), p. 178–191

[49] Goodman, M. ; Czelusniak, John ; Moore, G. W. ; Herrera, Romero A. ; Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. En: *Systematic Zoology* 28 (1979), p. 132–163

[50] Haft, Daniel H. ; Selengut, Jeremy D. ; White, Owen: The TIGRFAMs database of protein families. En: *Nucleic Acids Research* 31 (2003), Nr. 1, p. 371–373

[51] Haft, Daniel H. ; Selengut, Jeremy D. ; White, Owen: The TIGRFAMs database of protein families. En: *Nucleic acids research* 31 (2003), Nr. 1, p. 371–373

[52] Haussler, David Kulp D. ; Eeckman, Martin G Reese Frank H.: A generalized hidden Markov model for the recognition of human genes in DNA. En: *Proc. int. conf. on intelligent systems for molecular biology, st. louis*, 1996, p. 134–142

[53] van der Heijden, R. T. ; Snel, B. ; van Noort, V. ; Huynen, M. A.: Orthology prediction at scalable resolution by phylogenetic tree analysis. En: *BMC Bioinformatics* 8 (2007), Mar, p. 83

[54] Herrero, J. ; Muffato, M. ; Beal, K. ; Fitzgerald, S. ; Gordon, L. ; Pignatelli, M. ; Vilella, A. J. ; Searle, S. M. ; Amode, R. ; Brent, S. ; Spooner, W. ; Kulesha, E. ; Yates, A. ; Flicek, P.: Ensembl comparative genomics resources. En: *Database (Oxford)* 2016 (2016)

[55] Herrin, Brantley R. ; Cooper, Max D.: Alternative Adaptive Immunity in Jawless Vertebrates. En: *The Journal of Immunology* 185 (2010), Nr. 3, p. 1367–1374

[56] Huang, Shengfeng ; Chen, Zelin ; Yan, Xinyu ; Yu, Ting ; Huang, Guangrui ; Yan, Qingyu ; Pontarotti, Pierre A. ; Zhao, Hongchen ; Li, Jie ; Yang, Ping [u. a.]: Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. En: *Nature communications* 5 (2014)

[57] Hughes, Thomas A.: Regulation of gene expression by alternative untranslated regions. En: *Trends in Genetics* 22 (2006), Nr. 3, p. 119–122

[58] Kasahara, Masanori ; Suzuki, Takashi ; Du Pasquier, Louis: On the origins of the adaptive immune system: novel insights from invertebrates and cold-blooded vertebrates. En: *Trends in immunology* 25 (2004), Nr. 2, p. 105–111

[59] Kelley, James ; de Bono, Bernard ; Trowsdale, John: IRIS: a database surveying known human immune system genes. En: *Genomics* 85 (2005), Nr. 4, p. 503–511

[60] Kilpatrick, David C.: Animal lectins: a historical introduction and overview. En: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1572 (2002), Nr. 2–3, p. 187 – 197. – Animal Lectins. – ISSN 0304–4165

[61] Kolkman, Joost A. ; Stemmer, Willem P.: Directed evolution of proteins by exon shuffling. En: *Nature biotechnology* 19 (2001), Nr. 5, p. 423–428

[62] Korf, Ian: Gene finding in novel genomes. En: *BMC bioinformatics* 5 (2004), Nr. 1, p. 59

[63] Korf, Ian ; Yandell, Mark ; Bedell, Joseph: *BLAST*. Sebastopol, CA, USA : O'Reilly & Associates, Inc., 2003. – ISBN 0596002998

[64] Kristensen, David M. ; Wolf, Yuri I. ; Mushegian, Arcady R. ; Koonin, Eugene V.: Computational methods for Gene Orthology inference. En: *Briefings in bioinformatics* 12 (2011), Nr. 5, p. 379–391

[65] Kuzniar, A. ; van Ham, R. C. ; Pongor, S. ; Leunissen, J. A.: The quest for orthologs: finding the corresponding gene across genomes. En: *Trends Genet.* 24 (2008), Nov, Nr. 11, p. 539–551

[66] Kuzniar, A. ; van Ham, R. C. ; Pongor, S. ; Leunissen, J. A.: The quest for orthologs: finding the corresponding gene across genomes. En: *Trends Genet.* 24 (2008), Nov, Nr. 11, p. 539–551

[67] Lange, Christina ; Hemmrich, Georg ; Klostermeier, Ulrich C. ; López-Quintero, Javier A. ; Miller, David J. ; Rahn, Tasja ; Weiss, Yvonne ; Bosch, Thomas C. ; Rosenstiel, Philip: Defining the Origins of the NOD-Like Receptor System at the Base of Animal Evolution. En: *Molecular Biology and Evolution* 28 (2011), Nr. 5, p. 1687–1702

[68] Lechner, Marcus ; Findeiss, Sven ; Steiner, Lydia ; Marz, Manja ; Stadler, Peter F. ; Prohaska, Sonja J.: Proteinortho: detection of (co-) orthologs in large-scale analysis. En: *BMC bioinformatics* 12 (2011), Nr. 1, p. 1

[69] Lemaire, Patrick: Evolutionary crossroads in developmental biology: the tunicates. En: *Development* 138 (2011), Nr. 11, p. 2143–2152

[70] Li, L. ; Stoeckert, C. J. ; Roos, D. S.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. En: *Genome Res.* 13 (2003), Sep, Nr. 9, p. 2178–2189

[71] Li, Ruiqiang ; Zhu, Hongmei ; Ruan, Jue ; Qian, Wubin ; Fang, Xiaodong ; Shi, Zhongbin ; Li, Yingrui ; Li, Shengting ; Shan, Gao ; Kristiansen, Karsten [u. a.]: De novo assembly of human genomes with massively parallel short read sequencing. En: *Genome research* 20 (2010), Nr. 2, p. 265–272

[72] Litman, Gary W. ; Rast, Jonathan P. ; Fugmann, Sebastian D.: The origins of vertebrate adaptive immunity. En: *Nature Reviews Immunology* 10 (2010), Nr. 8, p. 543–553

[73] Litvack, Michael L. ; Palaniyar, Nades: Review: Soluble innate immune pattern-recognition proteins for clearing dying cells and cellular components: implications on exacerbating or resolving inflammation. En: *Innate Immunity* 16 (2010), Nr. 3, p. 191–200

[74] Loker, Eric S. ; Adema, Coen M. ; Zhang, Si-Ming ; Kepler, Thomas B.: Invertebrate immune systems–not homogeneous, not simple, not well understood. En: *Immunological reviews* 198 (2004), Nr. 1, p. 10–24

[75] Matthysse, Ann G. ; Deschet, Karine ; Williams, Melanie ; Marry, Mazz ; White, Alan R. ; Smith, William C.: A functional cellulose synthase from ascidian epidermis. En: *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004), Nr. 4, p. 986–991

[76] Medzhitov, Ruslan: Toll-like receptors and innate immunity. En: *Nat Rev Immunol* 1 (2001), Nov, Nr. 2, p. 135–145. – ISSN 1474–1733

[77] Merkeev, I. V. ; Novichkov, P. S. ; Mironov, A. A.: PHOG: a database of supergenomes built from proteome complements. En: *BMC Evol. Biol.* 6 (2006), Jun, p. 52

[78] Milinski, Manfred: The Major Histocompatibility Complex, Sexual Selection, and Mate Choice. En: *Annual Review of Ecology, Evolution, and Systematics* 37 (2006), Nr. 1, p. 159–186

[79] Miller, David J. ; Hemmrich, Georg ; Ball, Eldon E. ; Hayward, David C. ; Khalturin, Konstantin ; Funayama, Noriko ; Agata, Kiyokazu ; Bosch, Thomas C.: The innate immune repertoire in Cnidaria - ancestral complexity and stochastic gene loss. En: *Genome Biology* 8 (2007), Nr. 4, p. 1–13. – ISSN 1474–760X

[80] Mukhopadhyay, Subhankar ; Gordon, Siamon: The role of scavenger receptors in pathogen recognition and innate immunity. En: *Immunobiology* 209 (2004), Nr. 1–2, p. 39 – 49. – ISSN 0171–2985

[81] Muller, Werner ; Frank, Uri ; Teo, Regina ; Mokady, Ofer ; Guette, Christina ; Plickert, Gunter: Wnt signaling in hydroid development: ectopic heads and giant buds induced by GSK-3beta inhibitors. En: *International Journal of Developmental Biology* 51 (2002), Nr. 3, p. 211–220

[82] Nei, M. ; Rooney, A. P.: Concerted and birth-and-death evolution of multigene families. En: *Annu. Rev. Genet.* 39 (2005), p. 121–152

[83] In: Nonaka, Masaru ; Satake, Honoo: *Urochordate Immunity*. Boston, MA : Springer US, 2010, p. 302–310. – ISBN 978–1–4419–8059–5

[84] Ocampo, Iván D ; Cadavid, Luis F.: Mecanismos de respuesta inmune en cnidarios. En: *Red de Estudios del Mundo Marino, Remar* , p. 175

[85] Oda, Hiroki ; Takeichi, Masatoshi: Structural and functional diversity of cadherin at the adherens junction. En: *The Journal of Cell Biology* 193 (2011), Nr. 7, p. 1137–1146

[86] O'Neill, Luke A.: When Signaling Pathways Collide: Positive and Negative Regulation of Toll-like Receptor Signal Transduction. En: *Immunity* 29 (2008), Nr. 1, p. 12 – 20. – ISSN 1074–7613

[87] Page, R. D. ; Charleston, M. A.: From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. En: *Mol. Phylogenet. Evol.* 7 (1997), Apr, Nr. 2, p. 231–240

[88] Pålsson-McDermott, EM ; O'Neill, Luke A. *Building an immune system from nine domains*. 2007

[89] Patthy, László: Genome evolution and the evolution of exon-shuffling—a review. En: *Gene* 238 (1999), Nr. 1, p. 103–114

[90] Pearl, Frances ; Todd, Annabel ; Sillitoe, Ian ; Dibley, Mark ; Redfern, Oliver ; Lewis, Tony ; Bennett, Christopher ; Marsden, Russell ; Grant, Alistair ; Lee, David [u. a.]: The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. En: *Nucleic acids research* 33 (2005), Nr. suppl 1, p. D247–D251

[91] Pilipenko, EV ; Blinov, VM ; Chernov, BK ; Dmitrieva, TM ; Agol, VI: Conservation of the secondary structure elements of the 5'untranslated region of cardio-and aphthovirus RNAs. En: *Nucleic Acids Research* 17 (1989), Nr. 14, p. 5701–5711

[92] Potter, Simon C. ; Clarke, Laura ; Curwen, Val ; Keenan, Stephen ; Mongin, Emmanuel ; Searle, Stephen M. ; Stabenau, Arne ; Storey, Roy ; Clamp, Michele: The Ensembl analysis pipeline. En: *Genome research* 14 (2004), Nr. 5, p. 934–941

[93] Pryszcz, L. P. ; Huerta-Cepas, J. ; Gabaldon, T.: MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. En: *Nucleic Acids Res.* 39 (2011), Mar, Nr. 5, p. e32

[94] Putnam, Nicholas H. ; Butts, Thomas ; Ferrier, David E. ; Furlong, Rebecca F. ; Hellsten, Uffe ; Kawashima, Takeshi ; Robinson-Rechavi, Marc ; Shoguchi, Eiichi ; Terry, Astrid ; Yu, Jr-Kai [u. a.]: The amphioxus genome and the evolution of the chordate karyotype. En: *Nature* 453 (2008), Nr. 7198, p. 1064–1071

[95] Quesenberry, Michael S. ; Ahmed, Hafiz ; Elola, Maria T. ; O'Leary, Nuala ; Vasta, Gerardo R.: Diverse lectin repertoires in tunicates mediate broad recognition and effector innate immune responses. En: *Integrative and comparative biology* 43 (2003), Nr. 2, p. 323–330

[96] Remm, M. ; Storm, C. E. ; Sonnhammer, E. L.: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. En: *J. Mol. Biol.* 314 (2001), Dec, Nr. 5, p. 1041–1052

[97] Rentzsch, Fabian ; Fritzenwanker, Jens H. ; Scholz, Corinna B. ; Technau, Ulrich: FGF signalling controls formation of the apical sensory organ in the cnidarian Nematostella vectensis. En: *Development* 135 (2008), Nr. 10, p. 1761–1769. – ISSN 0950–1991

[98] Rosenstiel, P. ; Philipp, E. E. R. ; Schreiber, S. ; Bosch, T. C. G.: En: *Journal of Innate Immunity* 1 (2009), Nr. 4, p. 291–300. – ISSN 1662–811X

[99] Ruan, J. ; Li, H. ; Chen, Z. ; Coghlan, A. ; Coin, L. J. ; Guo, Y. ; Heriche, J. K. ; Hu, Y. ; Kristiansen, K. ; Li, R. ; Liu, T. ; Moses, A. ; Qin, J. ; Vang, S. ; Vilella, A. J. ; Ureta-Vidal, A. ; Bolund, L. ; Wang, J. ; Durbin, R.: TreeFam: 2008 Update. En: *Nucleic Acids Res.* 36 (2008), Jan, Nr. Database issue, p. D735–740

[100] Salamov, Asaf A. ; Solovyev, Victor V.: Ab initio gene finding in Drosophila genomic DNA. En: *Genome research* 10 (2000), Nr. 4, p. 516–522

[101] Satake, Honoo ; Sekiguchi, Toshio: Toll-Like Receptors of Deuterostome Invertebrates. En: *Frontiers in Immunology* 3 (2012), p. 34. – ISSN 1664–3224

[102] Schumann, Ralf R.: Old and new findings on lipopolysaccharide-binding protein: a soluble pattern-recognition molecule. En: *Biochemical Society Transactions* 39 (2011), Nr. 4, p. 989–993. – ISSN 0300–5127

[103] Schwarz, Ryan S. ; Hodes-Villamar, Linda ; Fitzpatrick, Kelly A. ; Fain, Matthew G. ; Hughes, Austin L. ; Cadavid, Luis F.: A gene family of putative immune recognition molecules in the hydroid Hydractinia. En: *Immunogenetics* 59 (2007), Nr. 3, p. 233–246. – ISSN 1432–1211

[104] Seo, Hee-Chan ; Kube, Michael ; Edvardsen, Rolf B. ; Jensen, Marit F. ; Beck, Alfred ; Spriet, Endy ; Gorsky, Gabriel ; Thompson, Eric M. ; Lehrach, Hans ; Reinhardt, Richard [u. a.]: Miniature genome in the marine chordate Oikopleura dioica. En: *Science* 294 (2001), Nr. 5551, p. 2506–2506

[105] Small, Kerrin S. ; Brudno, Michael ; Hill, Matthew M. ; Sidow, Arend: A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. En: *Genome biology* 8 (2007), Nr. 3, p. R41

[106] Stanke, M. ; Morgenstern, B.: AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. En: *Nucleic Acids Res.* 33 (2005), Jul, Nr. Web Server issue, p. W465–467

[107] Stanke, Mario ; Diekhans, Mark ; Baertsch, Robert ; Haussler, David: Using native and syntenically mapped cDNA alignments to improve de novo gene finding. En: *Bioinformatics* 24 (2008), Nr. 5, p. 637–644

[108] Tan, Kemin ; Duquette, Mark ; Liu, Jin-huan ; Dong, Yicheng ; Zhang, Rongguang ; Joachimiak, Andrzej ; Lawler, Jack ; Wang, Jia-huai: Crystal structure of the TSP-1 type 1 repeats: a novel layered fold and its biological implication. En: *The Journal of Cell Biology* 159 (2002), Nr. 2, p. 373–382

[109] Tatusov, Roman L. ; Galperin, Michael Y. ; Natale, Darren A. ; Koonin, Eugene V.: The COG database: a tool for genome-scale analysis of protein functions and evolution. En: *Nucleic acids research* 28 (2000), Nr. 1, p. 33–36

[110] Tatusova, Tatiana ; DiCuccio, Michael ; Badretdin, Azat ; Chetvernin, Vyacheslav ; Nawrocki, Eric P. ; Zaslavsky, Leonid ; Lomsadze, Alexandre ; Pruitt, Kim D. ; Borodovsky, Mark ; Ostell, James: NCBI prokaryotic genome annotation pipeline. En: *Nucleic Acids Research* (2016), p. gkw569

[111] Thomas, Paul D. ; Campbell, Michael J. ; Kejariwal, Anish ; Mi, Huaiyu ; Karlak, Brian ; Daverman, Robin ; Diemer, Karen ; Muruganujan, Anushya ; Narechania, Apurva: PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. En: *Genome Research* 13 (2003), Nr. 9, p. 2129–2141

[112] Thomas, Paul D. ; Campbell, Michael J. ; Kejariwal, Anish ; Mi, Huaiyu ; Karlak, Brian ; Daverman, Robin ; Diemer, Karen ; Muruganujan, Anushya ; Narechania, Apurva: PANTHER: a library of protein families and subfamilies indexed by function. En: *Genome research* 13 (2003), Nr. 9, p. 2129–2141

[113] Thorisson, Gudmundur A. ; Smith, Albert V. ; Krishnan, Lalitha ; Stein, Lincoln D.: The international HapMap project web site. En: *Genome research* 15 (2005), Nr. 11, p. 1592–1593

[114] Trachana, Kalliopi ; Larsson, Tomas A. ; Powell, Sean ; Chen, Wei-Hua ; Doerks, Tobias ; Muller, Jean ; Bork, Peer: Orthology prediction methods: a quality assessment using curated protein families. En: *Bioessays* 33 (2011), Nr. 10, p. 769–780

[115] Turvey, Stuart E. ; Broide, David H.: Innate immunity. En: *Journal of Allergy and Clinical Immunology* 125 (2010), Nr. 2, Supplement 2, p. S24 – S32. – 2010 Primer on Allergic and Immunologic Diseases. – ISSN 0091–6749

[116] Voskoboynik, Ayelet ; Neff, Norma F. ; Sahoo, Debashis ; Newman, Aaron M. ; Pushkarev, Dmitry ; Koh, Winston ; Passarelli, Benedetto ; Fan, H C. ; Mantalas, Gary L. ; Palmeri, Karla J. [u. a.]: The genome sequence of the colonial chordate, Botryllus schlosseri. En: *Elife* 2 (2013), p. e00569

[117] Walter, Lutz: Pas de deux: Natural Killer Receptors and MHC Class I Ligands in Primates. En: *Current genomics* 8 (2007), Nr. 1, p. 51–57

[118] Wheeler, Travis J. ; Eddy, Sean R.: nhmmer: DNA homology search with profile HMMs. En: *Bioinformatics* (2013), p. btt403

[119] Wilson, Derek ; Madera, Martin ; Vogel, Christine ; Chothia, Cyrus ; Gough, Julian: The SUPERFAMILY database in 2007: families and functions. En: *Nucleic Acids Research* 35 (2007), Nr. suppl 1, p. D308–D313

[120] Wu, Cathy H. ; Nikolskaya, Anastasia ; Huang, Hongzhan ; Yeh, Lai-Su L. ; Natale, Darren A. ; Vinayaka, CR ; Hu, Zhang-Zhi ; Mazumder, Raja ; Kumar, Sandeep ; Kourtesis, Panagiotis [u. a.]: PIRSF: family classification system at the Protein Information Resource. En: *Nucleic acids research* 32 (2004), Nr. suppl 1, p. D112–D114

[121] Yandell, M. ; Ence, D.: A beginner's guide to eukaryotic genome annotation. En: *Nat. Rev. Genet.* 13 (2012), Apr, Nr. 5, p. 329–342

[122] In: Zárate-Potes, Alejandra ; Cadavid, Luis F.: *Transcriptomics of the Immune System of Hydrozoan Hydractinia symbiolongicarpus Using High Throughput Sequencing Methods*. Cham : Springer International Publishing, 2014, p. 239–245. – ISBN 978–3–319–01568–2

[123] Zarbock, Alexander ; Ley, Klaus ; McEver, Rodger P. ; Hidalgo, Andrés: Leukocyte ligands for endothelial selectins: specialized glycoconjugates that mediate rolling and signaling under f low. En: *Blood* 118 (2011), Nr. 26, p. 6743–6751. – ISSN 0006–4971

[124] Zhang, Simo V. ; Zhuo, Luting ; Hahn, Matthew W.: AGOUTI: improving genome assembly and annotation using transcriptome data. En: *bioRxiv* (2015), p. 033019