UNIVERSIDAD NACIONAL DE COLOMBIA

Construcción de una plataforma computacional para la identificación de relaciones sinténicas entre subregiones genómicas asociadas al sistema inmune innato del organismo *D. vexillum* y otros tunicados

Ernesto Parra Rincón

Universidad Nacional de Colombia
Facultad de Ingeniería
Bogotá, Colombia
2020

Construcción de una plataforma computacional para la identificación de relaciones sinténicas entre subregiones genómicas asociadas al sistema inmune innato del organismo *D. vexillum* y otros tunicados

Ernesto Parra Rincón

Tesis de grado presentada(o) como requisito parcial para optar al título de:
Magister en Bioinformática

Directora:
Ph.D. Clara Isabel Bermúdez Santana

Línea de Investigación:
Genomica Comparativa
Grupo de Investigación:
Rnómica Teórica y Computacional

Universidad Nacional de Colombia
Facultad de Ingeniería)
Bogotá, Colombia
2020

A mi Familia

Agradecimientos

- A la profesora Clara Bermúdez de la Universidad Nacional de Colombia por su constante apoyo y confianza.
- Al Profesor Arjan Gittenberger del Instituto GiMaris y la Universidad de Leiden en Holanda por el proceso de colecta de la colonia de *D. vexillum* y el aislamiento genómico y transcriptómico.
- A Colciencias por el financiamiento otorgado para realizar este trabajo bajo el Proyecto No. 110165843196, contrato 571-2014.
- A los profesores Peter F. Stadler de la Universidad de Leipzig Alemania y a Federico Brown de la Universidad de Sao Paulo, Brasil por su apoyo y contrapartida para la ejecución conjunta del proyecto de Colciencias.
- Al Biólogo Candidato a PhD Cristian Velandia por su apoyo y acompañamiento en algunas etapas del proyecto.
- Se agradece al Sistema de Intercambio Alemán DAAD y a la Facultad de Ciencias por el fortalecimiento de equipos del Laboratorio de Biología Computacional de la Facultad de Ciencias de la Universidad Nacional de Colombia en donde se realizaron parte del proceso de computo.
- Se agradece al Laboratorio de Computo de Alto Rendimiento del Grupo de Bioinformática de la Universidad de Leipzig Alemania en donde se realizaron los computos de alta capacidad computacional requeridos para el ensamblaje y la comparación genómica.
- Al Ingeniero Jens Steuck del Grupo de Bioinformática de la Universidad de Leipzig Alemania por su apoyo en logística administrativa para el funcionamiento de equipos.

Resumen

En el presente trabajo se presentan los protocolos y resultados asociados a la identificación de relaciones sinténicas entre subregiones genómicas asociadas al sistema inmune innato de *Didemnum vexillum* y otros Tunicados. Para poder realizar estas comparaciones se requirió reensamblar el genoma combinando dos tecnologías de secuenciamiento: PacBio e Illumina. Adicionalmente, se incorporaron protocolos computacionales ajustados para superar dos problemas técnicos: primero el estado de fragmentación del genoma aislado y su secuenciamiento por tecnología PacBio y segundo la necesidad de corregir errores del secuenciamiento e incrementar profundidad del ensamblaje genómico usando lecturas de Illumina. El ensamblaje dihíbrido se hizo usando Celera Assembler Approach Version 8.3rc2. Posteriormente a las correciones propias de errores de secuenciamiento se usó SSPACE-Long. El ensamble final resultó en 517.5 Mb compuesto de 109.769 scaffolds con un N50 de 6.54 kb. Adicionalmente para poder identificar regiones genómicas con potencial de codificar genes asociados al sistema inmune innato se ensambló de novo el transcriptoma usando Trinity v2.4.0 produciendo un total de 90,938 transcritos putativos. Posteriormente se realizó la anotación estructural utilizando Maker en dos rondas de entrenamiento que incluyeron los transcritos ensamblados: La primera usó modelos de genes de la especie *Ciona intestinalis*(hoy conocida como *Ciona robusta*) y la posterior ronda incorporó modelos de genes entrenados con las secuencias de *Didemnum vexillum*. La anotación funcional se realizó usando términos del Gene Ontology(GO) obtenidos por homología a las bases de datos Uniref90 de Uniprot y PFAM.

Una vez obtenido el re-ensamblaje del genoma y el del transcriptoma de la especie se procedió a realizar la comparación genómica pareada entre el genoma de *D. vexillum* y los ensamblajes de genómicos de otras 4 especies de tunicados y de un cefalocordado utilizando el software SatSuma. Finalmente, sobre las regiones sinténicas obtenidas se asociaron aquellas que tienen genes asociados al sistema inmune según los términos GO. La información genómica y transcriptómica se visualiza en la plataforma base de datos GMOD ajustada para tal fín.

Palabras clave: Tunicados, Didemnum vexillum, Genómica, PacBio, Illumina, Sintenia, Sistema Inmune, GMOD.

Abstract

This works presents protocols and results associated to identify sintenic relationships between subgenomics regions associated with the innate immune system of the organism *Didemnum vexillum* and other Tunicates. In order to achive these comparisons, was necessary to re-assemble the genome of the organism by combining two sequencing technologies: PacBio and Illumina. Additionally, fitted computational protocols were incorporated to overcome two technical problems: first the state of fragmentation of the isolated genome and sequencing by PacBio technology and second the need to correct sequencing errors and increase genomic assembly deep by using Illumina reads. The di-hybrid assembly was run using Celera Assembler Approach Version 8.3rc2. Subsequently to corrections of sequencing errors, SSPACE-Long was used to genome scaffolding.

The final assembly resulted in 517.5 Mb comprised of 109,769 scaffolds with an N50 of 6.54 kb. Additionally, in order to identify genomic regions with the potential to encode genes associated with the innate immune system, the de novo transcriptome assembly was run using Trinity v2.4.0 producing a total of 90,938 putative transcripts. Subsequently, the structural annotation was performed using Maker using two training rounds. First round using reference models of the species *Ciona intestinalis*(today know as *Ciona robusta*) and the second consecutive round incorporates non-assembled transcriptome data from textit D. vexillum. Functional annotation was assigned with GO terms using Uniref90 from Uniprot and PFAM.

Once genome re-assembly and transcriptome assembly , pair-wise genome comparison was achieved between the genome of *D. vexillum* and four other tunicate species and with a cephalocordado by using the algorith SatSuma. Finally, those syntenic regions associated with GO terms of the immune system were detected. Genomic and transcriptomic information are shown on the platform GMOD database ajusted for this purpose.

Keywords: Tunicata, Didemnum vexillum, Genómics, PacBio, Illumina, Synteny, Inmune System, GMOD

Lista de Figuras

Lista de Tablas

Introducción

En el presente trabajo se compilan los esfuerzos computacionales requeridos para un análisis genómico comparativo, dirigido a proponer estrategias en el uso de métodos computacionales para resolver y superar los retos y las dificultades experimentales propias del secuenciamiento genómico de un organismo no modelo como lo es el tunicado *Didemnum vexillum*, con el fin de detectar subregiones genómicas asociadas a genes putativos relacionados con ontologias del sistema inmune innato que presenten conservación sinténica con otras especies de tunicados. *D. vexillum* es una especie que habita ambientes marinos y es considerado por la comunidad internacional como una especie altamente invasiva con potencial para modificar la estabilidad de los ecosistemas marinos. Este organismo es una especie omnívora, que en conjunto con su carácter de invasor, se espera, que hubiese desarrollado estrategias inmunológicas diversas y flexibles que le confirieron éxito en su proceso de adaptación a diversos ambientes. Evolutivamente la especie hace parte del *phylum* de los tunicados que es considerado el grupo hermano de los vertebrados y es modelo de estudio para la biología evolutiva del desarrollo. Estos estudios son dependientes de la expresión de genes en los diferentes estadios del desarrollo y del conocimiento de la estructura y funcionamiento del genoma. Sin embargo, la complejidad de los genomas reportados han indicado, rápidas tasas de mutación, rearreglos genómicos y un ejemplo claro de compactación del genoma como ocurre en la especie *Oikopleura dioca*. Esta complejidad genera retos para los estudios genómicos comparativos por dos motivos: por un lado la estructura propia del genoma y por otro, el estado actual de la información genómica, la cual en la mayoría de los tunicados se encuentra en estado borrador, lo cual dificulta aun más los estudios genómicos comparativos. De esta información, la excepción que ya se estructura a nivel de cromosoma es para las especies: *Ciona robusta*, antes clasificada como *Ciona intestinales* (nombre que se mantendrá a lo largo del texto) y *Botryllus schlosseri*.

Para este trabajo, los procedimientos computacionales fueron ajustados para integrar dos fuentes de tecnologías de secuenciamiento de ADN genómico(DNAg por su sigla en inglés), que junto a la información transcriptómica de la especie fueron incorporados en los estudios de genómica comparativa. Tanto el re-ensamblaje genómico como el ensamblaje del transcriptoma, fueron logrados para cumplir con los dos estándares mínimos requeridos en genómica comparativa y que son presentados en el primer capítulo: un genoma ensamblado al mejor nivel de resolución posible y el ensamblaje del transcriptoma. Por otro lado, en el capítulo segundo se presentan los resultados de la anotación estructural y funcional que serán incorporados en los estudios comparativos descritos al final del mismo; para cumplir

con los requisitos de aplicación de la genómica comparativa para el cálculo de las regiones sinténicas de comparaciones pareadas entre el genoma ensamblado de *D. vexillum* con los genomas de cada una de las especies a evaluar en este trabajo: comparaciones de *D. vexillum* con las especies *C. intestinalis*, *Ciona savigni*, *B. schlosseri*, *O. dioca* y del cefalocordado *Branchiostoma floridae* con la intercepción de bloques que contienen genes con ontologías asociadas al sistema inmune innato del fagosoma de la especie *C. intestinalis*. Finalmente en el capítulo tercero se usaron herramientas del proyecto *Generic Model Organism Database*(GMOD) para incorporar los resultados de los capítulos anteriores, en una base de datos que relaciona las secuencias genómicas, usando sus anotaciones de términos de ontología y en la que se pueden realizar consultas simple como indexaciones y búsquedas desde la WEB. Con esta aplicación se permite presentar los resultados dando un sentido biológico y nos posibilita hacer esta plataforma extensible a otros organismos.

Es importante resaltar que el análisis bioinformático aplicado a los estudios genómicos comparativos depende de los protocolos que no siempre pueden ser utilizados de forma estándar y, a su vez, son altamente dependientes de los resultados derivados de los experimentos por un lado y, altamente dependientes de la información de referencia de otros genomas disponibles por otro. En el caso del estudio del genoma de *D. vexillum* estos dos requerimientos demandaron el ajuste de los protocolos estándar para lograr los objetivos a cumplir.

Los resultados genómicos del secuenciamiento de *D. vexillum* obtenidos por tecnología PacBio con baja profundidad se originaron de un DNAg parcialmente degradado que limitó el uso de los protocolos en la plataforma SMRT Analysis suite–una plataforma concebida para datos generados desde secuenciamiento PacBio– y demandó el uso del ensamblador Celera Assembler. Por otro lado, la información genómica previamente publicada y obtenida por la tecnología Illumina, fue usada como información de referencia en la corrección de errores en las lecturas PacBio y en un proceso posterior de scaffolding conservativo fue procesada utilizando el programa SSPACE-Long. Finalmente se realizó con el protocolo Quiver de la SMRT Analysis suite v2.3.0 una fase de construcción de consensos de alta calidad, incluidas variantes tipo SNPs e INDELs.

La integridad de las secuencias codificantes en el ensamble genómico y el grado de fragmentación de los genes derivados del ensamblaje genómico fue evaluado usando el software BUSCO basado de una primera anotación estructural.

Por otro lado, se ensambló *de novo* el transcriptoma de la especie usando el software Trinity v2.4.0 produciendo un total de 90,938 transcritos putativos. La anotación estructural se realizó con Maker usando dos rondas de entrenamiento. La primera ronda de entrenamiento se realizó usando modelos de referencia de la especie *C. intestinalis* cuyos productos fueron usados para una segunda ronda de entrenamiento que incorpora los datos crudos del transcriptoma de *D. vexillum*. La anotación funcional se asignó con terminos del *Gene Ontology*(GO) y usando Uniref90 de Uniprot y PFAM.

Finalmente re-ensamblado el genoma y el transcriptoma de la especie se realizaron las comparaciones genómicas entre *D. vexillum* y los genomas de otras cuatro especies de tunicados

y del cefalocordado utilizando el software SatSuma. Una vez detectadas las regiones conservadas se interceptaron con aquellas que contiene genes asociados al sistema inmune según los términos GO reportados y el fagosoma disponible en *Kyoto Encyclopedia of Genes and Genomes*(KEGG).
En el tercer capítulo, la información genómica y transcriptómica es ajustada con herramientas del proyecto GMOD para ser incorporada en una base de datos construida con el esquema relacional Chado que, se presenta al investigador en un entorno web donde se usa el navegador genómico Gbrowse y la herramienta de búsquedas por homología Blast permitiendo acceso directo a todos los recursos y resultados de este trabajo.

Objetivo General

Construir una plataforma computacional que evidencie las relaciones sinténicas entre las sub-regiones genómicas asociadas al Sistema Inmune Innato del genoma de *Didemnum vexillum* y otros tunicados.

Objetivos Especificos

1. Re-secuenciar el genoma del tunicado *Didemnum vexillum* reportado por Velandia-Huerto et.al. (2016) con las secuencias genómicas nucleares obtenidas por la técnica de secuenciamiento PacBio.

2. Localizar las coordenadas de las subregiones genómicas asociadas al sistema inmune innato de los tunicados: *Didemnum vexillum, Ciona intestinalis, Ciona savignyi, Oikopleura dioica, Botryllus schlosseri* y el cefalocordado *Branchiostoma floridae*.

3. Establecer las relaciones sinténicas entre las subregiones genómicas asociadas al sistema inmune innato localizadas en los organismos: *Didemnum vexillum, Ciona intestinalis, Ciona savignyi, Oikopleura dioica, Botryllus schlosseri y el cefalocordado Branchiostoma floridae*.

4. Realizar la implementación del proyecto GMOD para construir la base de datos del organismo *Didemnum vexillum*, usando las secuencias genomicas nucleares, la anotación estructural, la anotacion funcional y la visualización de las regiones sinténicas asociadas al sistema inmune innato previamente identificado en este organismo y en otros tunicados.

1. Re-ensamblaje del Genoma de la especie *Didemnum vexillum*

En el presente capítulo se presentan los procedimientos y resultados para el re-ensamblaje del genoma del invertebrado marino *D. vexillum*. Estos resultados versan alrededor de un organismo que se ha informado como una especie altamente invasiva y de alto potencial para modificar la estabilidad de los ecosistemas marinos[Valentine *et al.*, 2006, Ordóñez *et al.*, 2015]. Es una especie omnívora e invasiva que probablemente obtuvo su éxito en términos de adaptación a diversos ambientes con el desarrollo de estrategias inmunológicas. Aunque esta hipótesis no ha sido probada, por carencia de datos experimentales para la anotación de genes asociados al sistema inmune, se ha encontrado que en estudios de poblaciones, esta especie puede vivir en aguas de temperaturas frías, en aguas tibias y en regiones subtropicales. Su ciclo de vida (crecimiento y reproducción), posee complejos patrones reproductivos en aguas templadas como regresiones de colonias en meses cálidos, teniendo su máxima abundancia en primavera [Ordóñez *et al.*, 2015].

Taxonómicamente, la especie hace parte del filo Cordados, subfilo Tunicados (Urocordados). Este subfilo evolutivamente es considerado como el grupo hermano más cercano a los vertebrados y debido a su ubicación evolutiva y a sus estadios del desarrollo, se incrementa cada vez mas la importancia de estudiar especies del subfilo, para usarlas como referencia en la biología evolutiva del desarrollo.

Desde el punto de vista genómico sólo han sido publicados siete genomas de tunicados, de los cuales la mayoría se encuentran ensamblados como borradores. El primer genoma fue secuenciado en el año 2001 y corresponde al de la especie *O. dioca*[Seo *et al.*, 2001], seguido por el genoma de *Ciona robusta* previamente conocido como *Ciona intestinalis Tipo A* en el 2002 [Dehal *et al.*, 2002] y el cual ha sido recientemente mejorado[Satou *et al.*, 2019]. En el año 2008 se publicó el genoma de la especie *C. savignyi* [Small *et al.*, 2007]; es importante mencionar que estas especies son de habitat solitario y sólo en el año 2013 se publicó el primer genoma de una especie con estilo de vida colonial, el genoma de la especie *B. schlosseri* [Voskoboynik *et al.*, 2013] y posteriormente el Grupo Rnomica Teórica y Computacional de la Universidad Nacional de Colombia[1], en conjunto con sus consorcios internacionales anotó y publicó en el año 2016 la primera versión borrador del genoma colonial de *D. vexillum*. Posterior a esta fecha se han publicado otros dos genomas, uno para la especie *Salpa thompsoni*[Jue *et al.*, 2016b] y otro para la especie *Botrylloides leachii* [Blanchoud *et al.*, 2018a].

[1] `http://ciencias.bogota.unal.edu.co/gruposdeinvestigacion/rnomica-teorica-y-computacional/`

Aunque muchos de los estudios de los tunicados son asociados al análisis de su biología, estos cada vez son mas dependientes del análisis de expresión de genes en diferentes estadios del desarrollo, por un lado, pero también son dependientes de los estudios genómicos por otro. Sin embargo, la complejidad esperada de los genomas de este grupo es alta como ha sido publicado y genera retos computacionales para el ensamblaje genómico. Por ejemplo, se han reportado rápidas tasas de mutación [Berná y Alvarez-Valin, 2014], rearreglos genómicos y un ejemplo claro de compactación del genoma como ocurre en la especie *O. dioca* [Seo *et al.*, 2001] así como inversiones recientemente descubiertas en el cromosoma 4 del género *Ciona* [Satou *et al.*, 2019] que fue previamente observado [Hill *et al.*, 2008]. Todos estos datos generaron retos para los estudios genómicos comparativos de este trabajo por un lado, y se añadió un segundo grado de complejidad, ya que la mayoría de los genomas de este grupo se encuentran en ensamblaje borrador y sólo para las especies *C. robusta* y *B.schlosseri* se encuentran los ensamblajes a nivel de cromosomas.

Es por esto que, con el interés de aportar al estudio del grupo tunicado, este capítulo responde al objetivo 1 de la propuesta de investigación: se presentan los resultados del re-ensamble genómico, incluidos los procedimientos computacionales para integrar dos fuentes de información genómica: secuencias largas de DNAg obtenidas del secuenciamiento PacBio [Eid *et al.*, 2009] y, secuencias cortas de DNAg producidas por tecnología Illumina utilizadas previamente para el primer ensamble del genoma de *D. vexillum*. El re-ensamble híbrido fue necesario porque compensa por un lado, los errores del secuenciamiento típico de PacBio, para el cual se reportan alta tasa de error técnico y por otro, la corrección que las secuencias Illumina aportan al proceso de ensamble. En este nuevo ensamble, el tamaño del N50 incrementó casi 8 veces en comparación la primera versión publicada y por ende permitió cumplir con el estándar mínimo requerido para realizar genómica comparativa por un lado y para poder hacer la anotación y predicción estructural. Sin embargo, pese a todos los esfuerzos de laboratorio, se presentó degradación del genoma previo al aislamiento, lo cual dificultó aun más el análisis computacional.

Se utilizó un ensamble *de novo* realizado con el software Celera Assembler [Myers *et al.*, 2000, Venter *et al.*, 2001] para poder manipular y ensamblar los productos de secuenciamiento de PacBio y además a los modelos de autocorrección de lecturas PacBio, se usaron los datos del secuenciamiento Illumina para hacer una segunda corrección. Posteriormente se dio la integración de información en un proceso de scaffolding conservativo utilizando el programa SSPACE-Long [Boetzer y Pirovano, 2014] con el uso final de QUIVER para una fase de pulimiento. Finalmente se evaluó la integridad de las secuencias codificantes y el grado de fragmentación de los genes derivados del ensamblaje genómico usando BUSCO[Simão *et al.*, 2015]. De los resultados de BUSCO se observa un comportamiento similar a lo observado para el genoma de la especie *S. thompsoni* en contraste al comportamiento observado para los demás tunicados usados para el análisis.

1.1. Fuente Biológica para el secuenciamiento PacBio de *D. vexillum*

Para el re-secuenciamiento del organismo *D. vexillum* se parte de dos muestras biológicas tomadas *in vivo* y rotuladas como Dvex2 y Dvex3; estas muestras fueron colectadas el 10 de junio de 2015 por el Profesor Arjan Gittenberger[2] durante un estudio de especies foráneas en lechos de ostras, en el lago marino Grevelingen(Holanda).
El secuenciamiento por técnica Pacbio fue realizado por los servicios de la Universidad de Washington en un instrumento PacBio RSII, usando química P6-C4 y utilizando los protocolos de preparación para plantillas con un tamaño de insertos de 5kbases[Biosciences, 2015c] y 20kbases[Biosciences, 2015b] para la muestra Dvex3 y protocolos de preparación de insertos de 10Kbases[Biosciences, 2015a] para la muestra Dvex2.
Las figuras **A-1** y **A-2** del anexo A presentan los electroferogramas del control de calidad a las muestras Dvex2 y Dvex3 realizados en la Universidad de Washington antes de ser preparadas y ligadas a los adaptores SMRTbell para PacBio: de este primer análisis experimental se cuantificó el grado de integridad del DNA de acuerdo a la escala DIN[Gassmann y McHoull, 2014] en $\leq$ 3,8, la distribución de tamaños la describe la tabla 1.1 y en ella se aprecia que $\sim 75\,\%$ de los fragmentos de DNAg de *D. vexillum* se agrupan por tamaños; de 1,5Kbp a 7Kbp y 7Kbp a 35Kpb. Un $42\,\%$ de los insertos se ubica en el primer grupo con un tamaño promedio de 3.6Kbp y con las mayores concentraciones: 9.72ng/ul para Dvex2 y 5.57ng/ul para Dvex3, el restante porcentaje de las librerías tienen longitud promedio de 15kbp, sin embargo, su concentración es la mas baja en las dos muestras: 3,71ng/ul para Dvex2 y 2,04ng/ul para Dvex3: considerando esta evidencia no se realizó el protocolo de normalización de tamaños antes del secuenciamiento PacBio[Wang *et al.*, 2019][3].

	From size(kb)	To size(kb)	Average size(kb)	CV size distribution	Conc. $[ng/\mu l]$
Dvex2	1.5	7	3.6	43.7 %	9.72
	7	35	15	32.3 %	3.71
Dvex3	1.5	7	3.7	42.6 %	5.57
	7	35	15.2	28.6 %	2.04

Tabla 1-1.: Distribución de tamaños del DNAg de *D. vexillum*, para las muestras Dvex2 y Dvex3 antes de ser preparadas para su secuenciamiento PacBio.

[2] https://www.gimaris.com/About-us/Principal-Research-Team/Arjan-Gittenberger

[3] https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-gDNA-\
Libraries-Using-the-SMRTbell-Express-Template-Preparation-Kit-2.0.pdf

1.2. Análisis Primarios del secuenciamiento PacBio del DNA genómico de *D. vexillum*

Los datos del secuenciamiento PacBio del DNAg de *D. vexillum*, fueron procesados por los scripts en python PacBioEDA v.2014-08-12[Skelly, 2014] para realizar los análisis primarios sobre las lecturas raw. La tabla B del anexo B relaciona los alias asignados a la extensa nomenclatura de nombres en Pacbio.

Productividad

Un paso crítico del secuenciamiento PacBio es la inmovilización del complejo(adaptadores, plantilla de DNA, polimerasa y primers) en el fondo de la Guía de Ondas Modo Cero(ZMW). La distribución óptima en la que un único complejo ocupa cada pozo ZMW de una celda SMRT de PacBio(∼150000 ZMW), sigue una distribución de Poisson; aproximadamente el 37 % (55,000 ZMW) de los pozos ZMWs podrían ser cargados con una única molécula de polimerasa[Biosciences, 2014](ver figura **1-1**).
Una medida de la carga o distribución del complejo en una celda SMRT es la **Productividad**, ella cuantifica el número de lecturas(campos de onda registrados) generados desde un ZMW en donde se ha inmovilizado el complejo: una productividad 0(**P0**) cuantifica los pozos ZMW vacios(sin polimerasa), una productividad 1(**P1**) define los pozos ZMW que son productivos y secuenciados (con polimerasa) y Productividad 2(**P2**) determina los pozos que no están vacíos(P0) o no son Productivos(P1)en donde las lecturas no son usables. La tabla 1.2 muestra los resultados de los análisis primarios productividad, valores de lecturas, sublecturas y bases obtenidas para las regiones de alta calidad(HQ)sobre los datos raw PacBio del organismo *D. vexillum* con el software PacBioEDA:

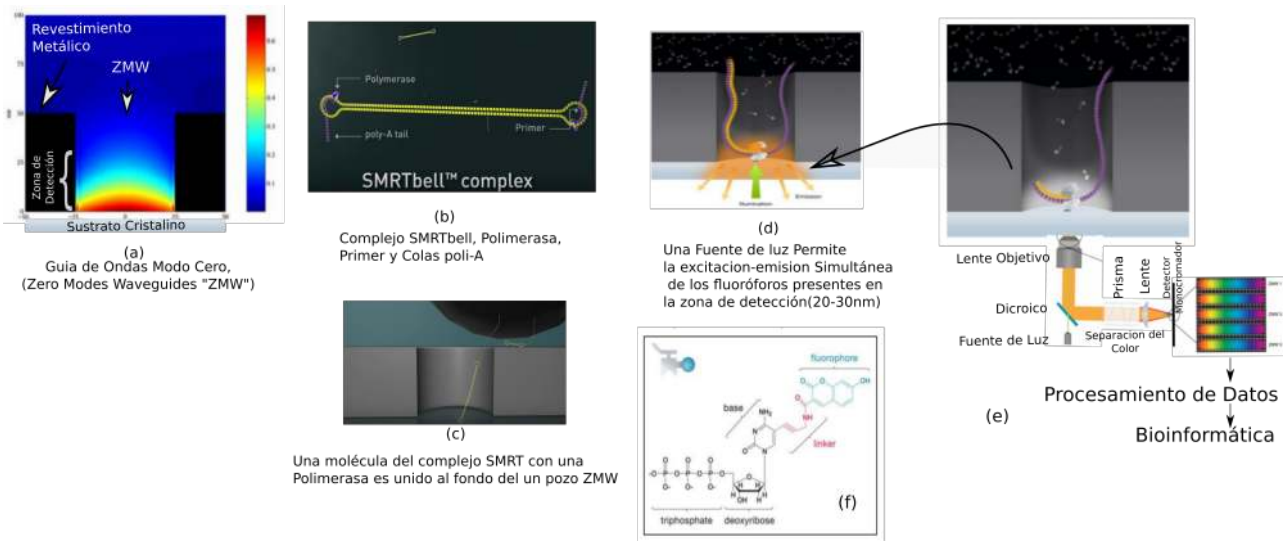


Figura 1-1.: Esquema General Técnica de secuenciamiento PacBio

(a) Diagrama de Guía de Ondas Modo Cero(ZMW), (b) Diagrams de complejo con adaptadores SMRTbell que ciclan un fragmento de DNA, a la que también se ligan Primers y una molécula de Polimerasa. (c) Una molécula del complejo (b) es unido al fondo de un pozo ZMW usando esferas paramagnéticas. (d) La luz emitida permite la excitación - emisión simultánea de los fluoróforos en la zona de detección(20-30nm).(e) la luz emitida por los flouróforos es colectada a través de un lente objetivo de gran apertura, luego la luz pasa a través de un prisma que dispersa los elementos que fueron deflectado por el cristal dicrónico, de acuerdo a su color. Se crea un patrón individual para cada ZMW, esto permite identificar el tipo de base que produce la señal y Un detector de alta sensibilidad es usado para identificar y discriminar pulsos para cada pozo ZMW. Las imágenes individuales fueron tomadas y adaptadas de Pacific “Biosciences Technology Backgrounder”

En esta etapa se descartaron las lecturas Cell_12, Cell_13 y Cell_14 de los métodos posteriores(Sombreado en la tabla1.2) debido a que el análisis primario evidencia un elevado porcentaje de pozos sin polimerasa(*P0*) y un bajo numero de bases en la region Alta Calidad(HQ).

			Productive	% Productivity			HQ			Average
Sample	Name	Protocol	ZMWs	P0	P1	P2	Reads	Subreads	Megabases	Kbases
Dvex2	Cell_1	10kb	150292	59.8	34.4	5.6	51791	110254	352.4	3.1
	Cell_2			32.8	54.3	12.7	81728	234344	696.2	2.9
	Cell_3			37.8	51.5	10.5	77536	223170	643.2	2.8
	Cell_4			33.5	51.7	14.7	77712	230168	669.4	2.9
Dvex3	Cell_5	5kb	150292	9.5	77.1	13.3	115915	705692	1782.6	2.5
	Cell_6			30	60	8.5	90986	514254	1104.2	2.1
	Cell_7			31	59.7	9.2	89803	508765	1106.8	2.1
	Cell_8			36.6	54.4	8.9	81787	458059	979.5	2.1
	Cell_9			6.2	71.8	21.9	107915	737216	1723.1	2.3
	Cell_10			5.1	72.8	21.9	109540	740399	1716.1	2.3
Dvex3	Cell_11	20kb	150292	20.3	66.6	13	100161	601263	1291.5	2.1
	Cell_12	20kb	150292	98.1	1.6	0.02	2420	4454	8.2	1.8
	Cell_13			97.9	1.7	0.2	2649	6427	12.1	1.8
	Cell_14			75	22	2.7	33433	40809	80.3	1.9

Tabla 1-2.: Análisis primarios sobre los datos raw PacBio de D. vexillum con el software PacBioEDA v.2014-08-1

1.3. Análisis Secundarios de los datos del secuenciamiento PacBio del DNAg *D. vexillum*

Los análisis secundarios para los datos producto del secuenciamiento PacBio del DNA genómico de *D. vexillum* se enfocaron en obtener lecturas libres de los errores inherentes a la técnica.

Fuente de error en las lecturas Pacbio de *D. vexillum*

Debido a que la identificación del oligo se basa en el tiempo de permanencia de la señal del fluoróforo unido al nucleótido incorporado en la réplica de la plantilla de DNA, cuando un

nucleótido fluoromarcado ingresa a la zona de detección durante un periodo de tiempo mas largo que el promedio pero no es incorporado, la técnica PacBio puede determinar incorrectamente que una base ha sido adicionada(errores tipo Inserción), algunas otras fracciones de nucleótidos a incorporar por la polimerasa evaden el etiquetado con el fluorocromo, al igual que posteriores etapas de purificación por lo que, la técnica puede no registrar la incorporación de la base cuando elonga la cadena réplica(errores tipo deleción)[Mardis, 2013]. Aunque también pueden existir errores debidos a la fusión de pulsos, como cuando se sintetizan cadenas de DNA con secuencias largas de homopolímeros, los errores predominantes son tipo InDels con inserciones de ∼12 % y deleciones ∼2 % [Contreras *et al.*, 2016]

1.3.1. Corrección de errores de las Lecturas PacBio *D. vexillum*

La alta tasa de error en las lecturas raw de Pacbio limita su uso directo por lo que, se han desarrollado métodos de corrección que caen en dos categorías: autocorrección y corrección híbrida[Shuhua *et al.*, 2019].

1.3.2. Corrección de errores usando una estrategia de Autocorrección

El software SMRT Analysis v2.3.0(Pacific Biosciences, CA) toma ventaja del error estocástico en la técnica PacBio y usando las lecturas raw de *D. vexillum* el módulo *RS_ReadsOfInsert* filtró las sublecturas CCS con ciclos completos ≥ 2 para luego superponerlas y generar consensos[Tedersoo *et al.*, 2018] de 220,514 lecturas CCS de una fiabilidad de $\geq 99\,\%$ y N50 2,1Kbp que corresponden a 450Mbp de lecturas de la mas alta calidad con una profundidad media de ∼13.8(ver tabla 1.3.2).

Paralelamente el mismo software uso el módulo de autocorrección *RS_PreAssembler*, que filtro secuencias de longitud $\geq$ 250bp con una calidad $RQ \geq 0{,}83$ para generar 823,758 *lecturas pre-ensambladas* de un tamaño de 1.4Gpb y un N50 1,8Kbp.

1.3.3. Corrección de errores usando una estrategia Híbrida

Usar únicamente una estrategia de autocorrección no fue factible con las lecturas Pacbio de *D. vexillum* debido a su baja profundidad: en una estrategia de corrección híbrida se usaron los datos de lecturas cortas de illumina de mayor cobertura y precisión para correguir la alta tasa de error y baja profundidad de las lecturas PacBio.

	Reads	Mbases	Average kbases	Accuracy %	Average No.Passes
Cell_1	5998	12.6	2.1	99.7	17.5
Cell_2	2613	7	2.6	99.7	15.9
Cell_3	30640	61	1.9	99.6	11.9
Cell_4	31743	63.1	1.9	99.6	11.9
Cell_5	6254	13.3	2.1	99.7	17.6
Cell_6	28709	57.8	2	99.6	10.9
Cell_7	23930	50.6	2.1	99.7	12.5
Cell_8	6502	14.9	2.3	99.7	17.3
Cell_9	31579	63.4	2	99.7	12.1
Cell_10	26799	56.1	2	99.6	12.3
Cell_11	25747	54.7	2.1	99.6	12.4
Average			2.1	99.6	13.8
$\sum$	220514	$\sim 453 Mbases$			

Tabla 1-3.: Métricas de calidad de Lecturas CCS de once celdas SMRT de PacBio con las muestras Dvex2 y Dvex3 del organismo *D. vexillum*. obtenidas con el módulo *RS_ReadsOfInsert* de SMRT Analysis v2.3.0(Pacific Biosciences, CA con ciclos completos ≥ 2.

1.3.4. Fuente biológica y datos de partida para la corrección de errores usando una estrategia híbrida

Lecturas cortas Illumina del DNAg de *D. vexillum*

Nuestro grupo de investigación describío el procedimiento seguido para colectar una colonia de *D. vexillum* en el islote Hompelvoet (Grevelingen, Holanda) en el año de 2009 [Velandia *et al.*, 2016]. Del DNA total de esta colonia se realizaron dos secuenciamientos de librerías pareadas(Run1 y Run3) en un equipo Illumina GAII: Las lecturas raw obtenidas(Anexo C, tabla C.1)con un posterior tratamiento serán los primeros datos de partida usados en la corrección de errores PacBio siguiendo una estrategia híbrida.

Pre-tratamiento de las Lecturas cortas Illumina del DNAg de *D. vexillum*

Los resultados de la evaluación de calidad a las lecturas raw Illumina del DNAg de *D. vexillum* usando el software FastQC,v.0.11.4[2010], puede verse en el anexo D.1; para complementar este análisis se usaron métodos k-mers de la suite BBtools[Bushnell, 2016][Bushnell *et al.*, 2017] que identificaron secuencias de dimeros de PCR y adaptadores ligados a las lecturas sentido

y antisentido derivados bien caracterizado genoma PhiX que es usado como control en los experimentos illumina[Illumina, 2017]. Estos contaminantes fueron insertos en la etapa del secuenciamiento Illumina y se extienden en diferentes posiciones dentro de las lecturas, como los muestra en resaltado de la figura **1-2**.

RUN1

>Read1_adapter
AGATCGGAAGAGCGGNTCAGCAGGAATGCCGAGACCG
La secuencia hace parte de pcr_dimer:
...ACTCTTTCCCTACACGACGCTCTTCCGATCT AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG ATCTCGTATGCCG...
La secuencia hace parte de PhiX_read1_adapter:
AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG ATCTCGTATGCCGTCTTCTGCTTGAAA

>Read2_adapter
AGATCGGAAGAGCGTCGTGTAGGGNAAGAGNGNAGAT
La secuencia hace parte de PhiX_read2_adapter:
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGAT CTCGGTGGTCGCCGTATCATTAAAAAA

RUN3

>Read1_adapter
AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATNTCGTATGCCGTCTTCTGCTTG
La secuencia hace parte de pcr_dimer:
AAT...TACACGACGCTCTTCCGATCT AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG
La secuencia hace parte de PhiX_read1_adapter:
AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG AAA

>Read2_adapter
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
La secuencia hace parte de PhiX_read2_adapter\cmidrule(l){3-12}
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT AAAAAA

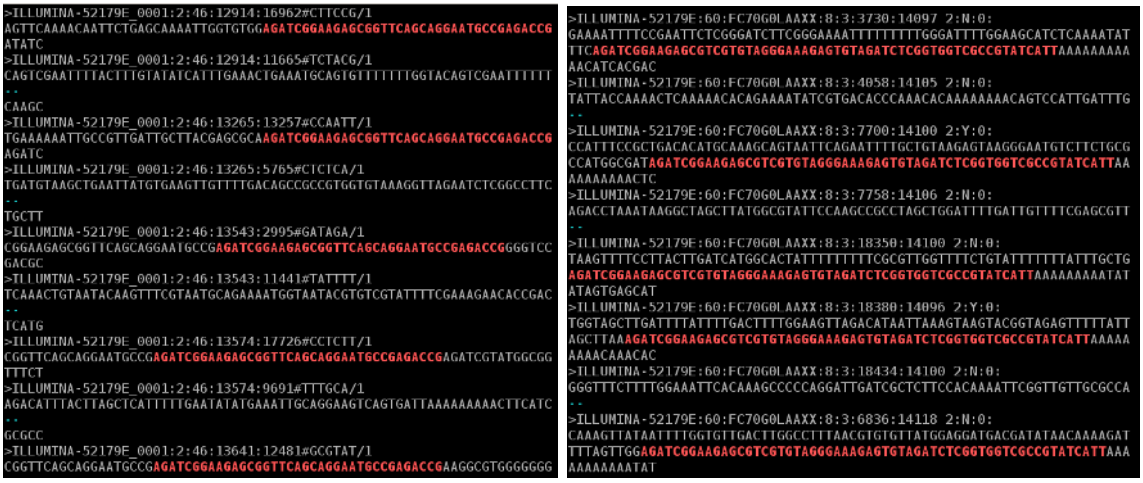


Figura 1-2.: Contaminantes identificados en las lecturas illumina del DNAg de *D. vexillum*

Tratamiento a los datos Illumina del DNAg de *D. vexillum*

La figura **1-3** relaciona la profundidad obtenida con un genoma referencia de 550Mbases sobre las lecturas illumina(tabla E del anexo E) **1.** sin ningún tratamiento. **2.** Luego de realizar recorte de los adaptadores identificados(PhiX y Dimeros de PCR), **3.** Recorte de los adaptadores identificados y exclusión de lecturas con calidad phRed $\geq$10, **4.** Recorte de los adaptadores identificados y recorte a las lecturas con calidad phRed $\geq$30.
Sobre las lecturas raw de Illumina, con el fin de mantener una profundidad moderada ($\sim 30X$), la herramienta bbduk de BBTools[Bushnell, 2016] realizó un enmascaramiento de los adaptadores identificados en las lecturas raw Illumina del organismo *D. vexillum* generando aproximadamente 15,2$Gbases$(1.3.4) que fueron usadas como entrada en la etapa de corrección hibrida de errores de las lecturas PacBio de nuestro organismo.

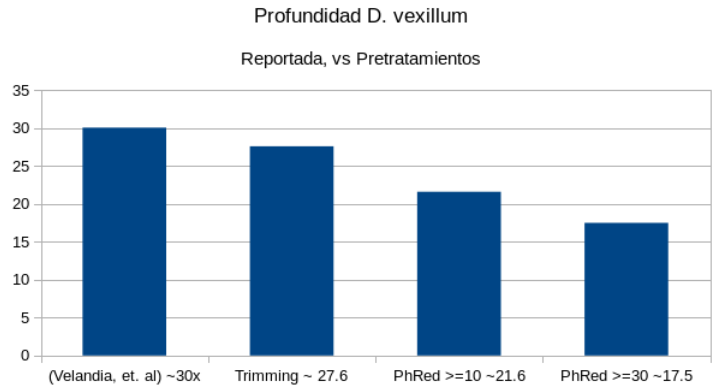


Figura 1-3.: Profundidad de secuenciamiento Illumina con un genoma referencia de 550Mbases.

		GC	GC stdev	A	C	G	T	Millones Lecturas	N90	N50	Longitud Total
Run1	Forward	0.36	0.09	0.32	0.17	0.19	0.3092	35	65	65	2,34Gbases
	Reverse	0.36	0.10	0.30	0.18	0.18	0.32	35,1	65	65	2,34Gbases
		GC	GC stdev	A	C	G	T	Millones Lecturas	N90	N50	Longitud Total
Run3	Forward	0.38	0.09	0.31	0.19	0.19	0.29	37.4	151	151	5.27Gbases
	Reverse	0.38	0.08	0.30	0.19	0.18	0.31	37,4	151	151	5.26Gbases

Tabla 1-4.: Métricas de calidad de Lecturas raw Illumina luego de Identificar y enmascarar los adaptadores insertos en la etapa de secuenciamiento.

Generación de las Lecturas Pacbio de DNAg de *D. vexillum* a correguir

La mayor población de sublecturas raw PacBio de *D. vexillum* a ser correguidas se obtuvieron usando el software `dextract`[Myers, 2015] que, filtro $5{,}1Giga - sublecturas$ de longitud $\geq 120bp$ y calidad $RQ \geq 0{,}75$ con un tamaño de 12.35 Gbases y N50 de 2.3kb. La frecuencia y calidad(RQ) en función de su longitud se ilustra en la figura **1-4**.

Corrección de errores de las Lecturas Pacbio de *D. vexillum* usando una estrategia híbrida

La sublectura PacBio obtenidas en el paso anterior1.3.4 fueron corregidas usando los datos de Illumina libres de errores(seccion 1.3.4) y las Sublecturas autocorreguidas CCS (seccion 1.3.1) por medio de la herramienta `proovread-2.13.13`[Förster *et al.*, 2014]. Como resultado se obtuvo un primer conjunto de lecturas PacBio correguidas con $776,295 Sublecturas$ de$N50{=}94kbases$ (untrimed) correspondientes a $2{,}7Gbases$ y otro conjunto de $288,198$ lecturas de mayor calidad con $N50 = 1{,}7kbp$ (trimmed) correspondientes a $391Mbp$.

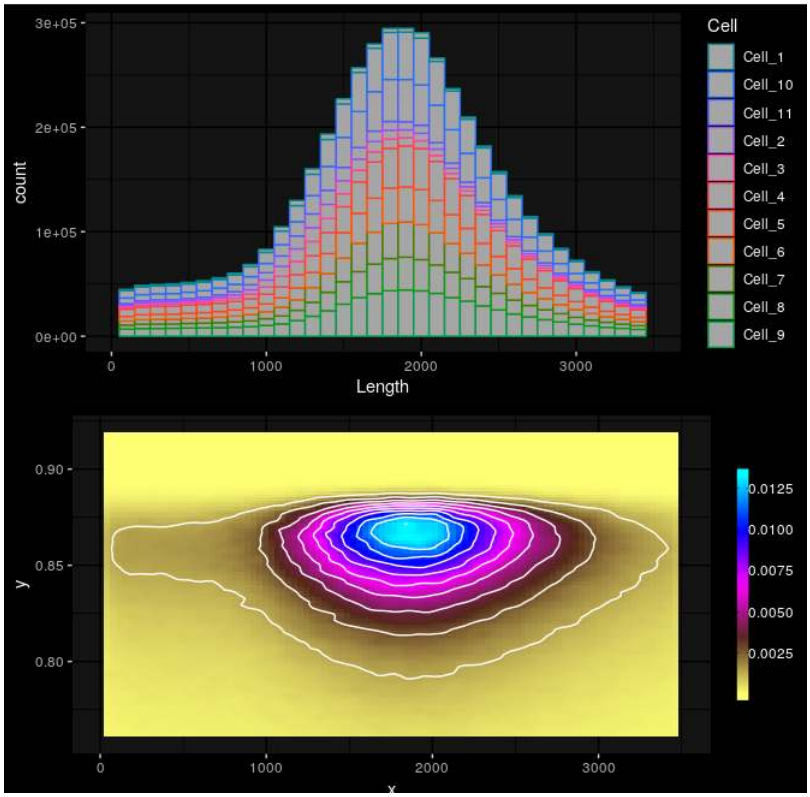


Figura 1-4.: Histograma con la distribucion de Sublecturas Raw Pacbio(Count) diferenciando las celdas SMRT de PacBio secuenciadas por color y su relacion con la longitud(x); El heatmap presenta una distribución de densidades(y) con respecto a la longitud de las lecturas(x)

1.3.5. Ensamble Genómico

Datos de entrada para el ensamble genomico de *D. vexillum*

La tabla **1-5** muestra 4,94$Gbases$ de sublecturas correguidas PacBio del organismo *D. vexillum* que fueron usadas como entrada al ensamble genómico.

Tabla 1-5.: Metricas de lecturas correguidas para el ensamble de D. vexillum

	Método de Corrección	Estrategia	Software	Reads	N50	Size
1	Corrección Hibrida	Alineamientos-Consenso	Proovread	776,295	3.94kbp	2.7Gp
2	Corrección Hibrida	Alineamientos-Consenso	Proovread	288,198	1.7kbp	391Mbp
3	Autocorrección	Solapamientos-Consenso	HGAP	823,758	1.8Kbp	1.4Gpb
4	Autocorrección	Solapamientos-Consenso	HGAP CCS	220,514	2.1Kbp	450 Mbp

Un ensamble *de novo* fue realizado con el software Celera Assembler[Myers *et al.*, 2000], version 8.3rc2, ajustado para no inducir rutas alternativas por superposiciones de los grafos(utgBubblePopping=0) y con tasas de error utgErrorRate=0.12, utgErrorLimit=2.5, ovlErrorRate=0.15, cnsErrorRate=0.15, cgwErrorRate=0.15 con un kmer=17. El procedimiento general seguido se describe en la figura **G-1**. La primera versión produjo un ensamble con 130,707 contigs, de un tamaño de 566,4Mpb con $N50 = 5{,}97kb$ and $GC = 36\,\%$.

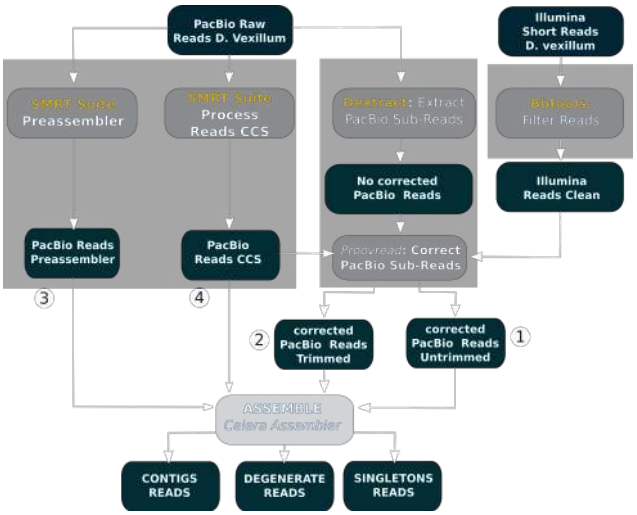


Figura 1-5.: Procedimiento general de ensamble de *D. vexillum* con lecturas Pacbio Correguidas(1, 2 ,3 y 4 corresponden a los datos de entrada en la tabla **1-5**)

La redundancia en el ensamble se redujo usando el software fasta2homozygous [Pryszcz y Gabaldón, 2016] que excluyó 16839 contigs de un tamaño $\leq 500bp$ y una identidad $\geq 95\,\%$ que corresponde a 47,4$Mbases$; los restantes 519Mbp fueron usados en la siguiente fase de scaffolding.

1.3.6. Scaffolding del ensamble genómico de *D. vexillum* con lecturas largas

Los métodos computacionales usados hasta aqui tienen dos etapas: la primera de ellas se realizó con Celera asembler que lectura a lectura construyó contigs y la segunda etapa del ensamble que fue el scaffolding, que normalmente hace uso de la información obtenida del secuenciamiento con finales pareados que logra anclar el ordenamiento y la orientación de los contigs. Sin embargo, los resultados finales son dependientes de la calidad y longitud del inserto en las lecturas[Qin *et al.*, 2019] por lo que el scaffolding del ensamble genómico del DNA de *D. vexillum*, en una primera etapa obtuvo lecturas largas Pacbio que se corrigieron usando grafos de Bruijn que fueron construidos usando las lecturas y la redundancia de las secuencias cortas Illumina.

Generación de las Lecturas Largas Pacbio de *D. vexillum* de alta calidad

Desde los datos raw de Pacbio se obtuvieron los consensos CCS desde la misma plantilla SMRTbell(Figura **F-1** en Anexo F) usando el software Unanimity[PacBio, 2015] v.3.0(–minPasses 3 –minPredictedAccuracy 0.9) generando 462447 sublecturas CCS en 985 Mbases con un N50=2.1kb: estas lecturas PacBio se sometieron a una corrección de error con el programa Lordec[Salmela y Rivals, 2014] que uso grafos de Bruijn construidos con las lecturas de Illumina libres de error con k-mers de longitud 13,15, 17, 19, 21, 31, 41 y 51.

En un siguiente paso, 11,428,46 sublecturas CCS con 7.2Gbases y N50=1.2Kbases libres de error fueron alineadas a si mismas con el software daligner[Myers, 2016](Tasa de correlación promedio de -e0.95) que produjo los alineamientos locales que fueron usados como entrada al software daccord[Tischler y Myers, 2017] que genero 1,9Giga lecturas consensos CCS con N50=1.5kbases en 2,3Gbases que Junto a los contigs del ensamble genomico fueron la entrada al sofware SSPACE-Long[Boetzer y Pirovano, 2014] para realizar un sccafolfing conservativo.

En esta etapa se generaron 110,006 scaffolds de un tamaño de Megabases con un N50=6.54Kb y contenido GC=36 con una desviación estándar de 0.0246.

1.3.7. Pulido del Ensamble

En esta etapa se vinculan las secuencias de scaffolds a valores de calidad(phRed), se identifican variables(SNP) y regiones en las que Pacbio puede presentar errores INDELS. El algoritmo QUIVER v.2.1[20] fue usado desde el software SMRT Analysis suite v2.3.0 para obtener siguiendo un protocolo de re-secuenciamiento (Modulo BAM Resequencing Beta.1) que tomó como referencia las secuencias generados en la etapa de scaffolding de *D. vexillum*.

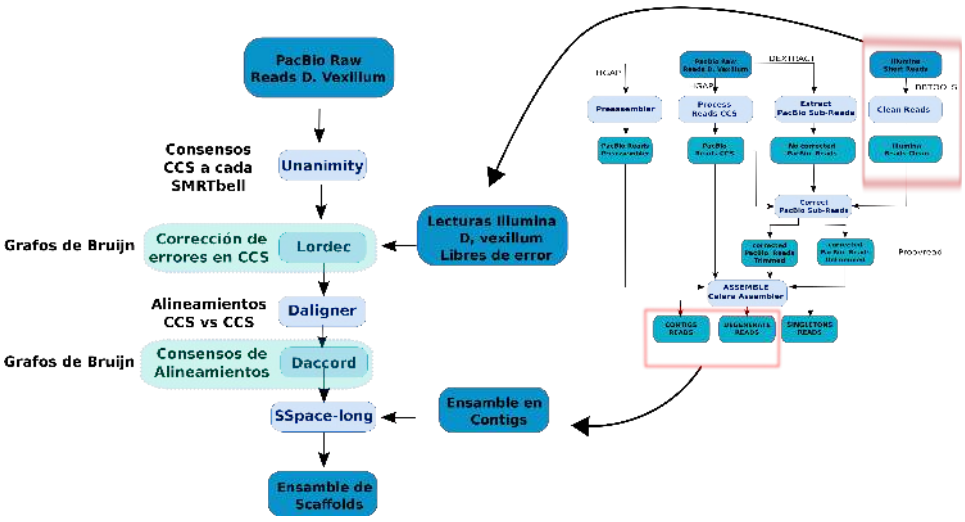


Figura 1-6.: Procedimiento del Scaffolding del ensamble en contigs de *D. vexillum*

1.3.8. Evaluación de la calidad del ensamblaje Genomico *D. vexillum*.

BUSCO [Simão *et al.*, 2015] evalúo la integridad de las secuencias codificantes y el grado de fragmentación de los genes en el ensamblaje genómico usando un conjunto de datos integrado por 978 ortólogos diferentes originados desde 65 especies del phylum metazoario metazoa_odb9(2016-02-13) el cual es usado como un gold standard de calidad del ensamblaje genómico. En el genoma de *D. vexillum*, un 50.8 % de los ortólogos se identificaron de manera completa(C); 48.6 % son genes únicos(S) y un 2.2 % son genes duplicados(D). 194 secuencias ortólogas fueron parcialmente identificadas y las restantes 287 secuencias evaluadas no hallaron ningún tipo de acierto.

En comparación con los genomas reportados de otros tunicados, el estado actual del ensamble genómico de *D. vexillum* en términos de integridad es similar al ensamble de *Salpa thompsoni*. En términos generales, la mayoría de los tunicados reportados muestran una similaridad $\geq 75{,}4\,\%$ en la identificación de los ortólogos completos de BUSCO**1-7**.

Finalmente se obtuvo un genoma de un tamaño de 517Mb distribuidos en 109.769 Scaffolds con un L50 de 25.281 y un N50 de 6.54. Los resultados comparativos con la versión previamente ensamblada se presentan en la tabla 2.2.3. Se observa que el nuevo genoma ensamblado se encuentra a nivel de scaffolds y no de contigs como en la versión anterior y el N50 es 8 veces mejor aproximadamente al obtenido previamente, lo cual favoreció los procesos de anotación de proteínas en los pasos siguientes de este trabajo.

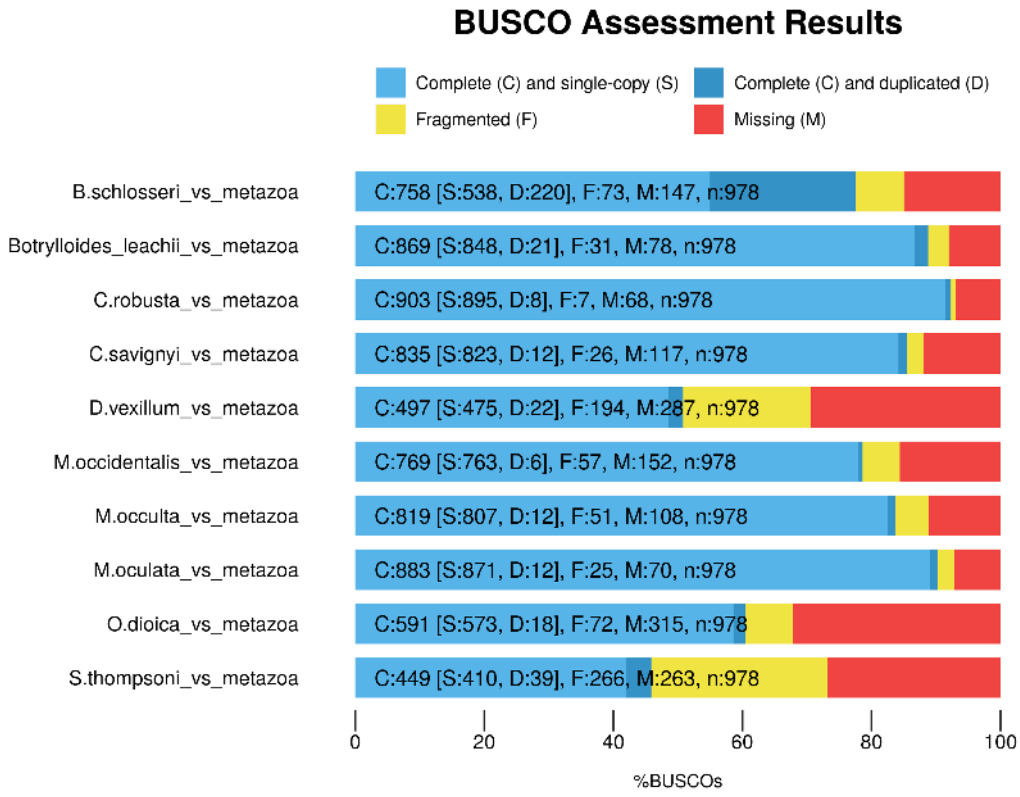


Figura 1-7.: Resultados de la evaluación del ensamble genómico de *D. vexillum* con la herramienta BUSCO

Ensamble	Tamaño Total(Kb)	Número de Contigs/Scaffols	L50 (Kb)	N50 (Kb)	Contenido GC±CV	Número de Genes	Número de Proteinas
Velandia-Huerto et.al.	542.259	882.106 Contigs	152.090	0.918	0.366±0.063	–	–
Este Trabajo	517.553	109.769 Scaffolds	25.281	6.54	0.362±0.024	62.194	64.424

Tabla 1-6.: Tabla comparativa entre el ensamble reportado por [Velandia *et al.*, 2016] en 2016 y el ensamble de *D. vexillum* producido de este trabajo

1.3.9. Ensamble Transcriptómico

Fuente biológica del RNA de *D. vexillum*

En 25 de Marzo de 2009 en el muelle sur del islote Hompelvoet (Grevelingen, Holanda) el Profesor Arjan Gittenberger[4] colectó una muestra de una colonia de *D. vexillum* de la cual se extrajo el RNA requerido para los análisis transcriptómicos. El secuenciamiento del RNA se realizó en un equipo GAIIX previa preparación de las librerías con finales pareados y un tamaño de inserto de 76 nucléotidos. Un total de 3.2Gpb de mRNA, distribuidas en 65.3 millones de lecturas con un $N50 = 50$ son los datos de partida en este trabajo.

Ensamble transcriptómico de *D. vexillum*

De los datos obtenidos del secuenciamiento del RNA por tecnología Illumina, se excluyeron las secuencias de adaptadores insertos por la técnica en el procedimiento de *trimming*, se usó los métodos k-mers de la suite BBtools[Bushnell, 2016] generando 55 millones de lecturas de RNA con finales pareados, de un $N50$ de $50bp$ y con un valor de calidad phRed $\geq$30. Estas lecturas se alinearon al genoma usando el programa GSNAP(Version 2019-06-10) [Wu *et al.*, 2016]. Los alineamientos obtenidos y las lecturas filtradas sirvieron de entrada al software Trinity v2.4.0 [Grabherr *et al.*, 2011] con los que realizó los procedimientos *ensamble de transcritos de novo* y *ensamble de transcritos guiado*, generando 90938 transcritos de *D. vexillum* en 39Mpb de un $N50 = 459$bp.
Sobre los transcritos se usó el software Transdecoder [Haas *et al.*, 2016] que identificó el marco abierto de lectura mas largo en cada una de las secuencias y desde ellas generó los candidatos a proteinas de *D. vexillum*. Las secuencias candidatas a proteína fueron evaluadas por su homología con las proteínas reportadas de los tunicados *C. intestinalis*, *B. schlosseri*, *C. savignyi*, *O. dioica* y el cefalocordado *B. floridae*. La figura **1-8** resume los resultados obtenidos y su homología usando Blast[Altschul *et al.*, 1990] con un e-value E-25. Los transcritos y las secuencias candidatas a proteína de *D. vexillum* fueron usadas en la etapa de anotación estructural con Maker presentada en el siguiente capítulo. Nótese que

[4] `https://www.gimaris.com/About-us/Principal-Research-Team/Arjan-Gittenberger`

entre un intervalo de 5184 a 6019 se encuentran distribuidos los homólogos resultantes de las comparaciones realizadas.

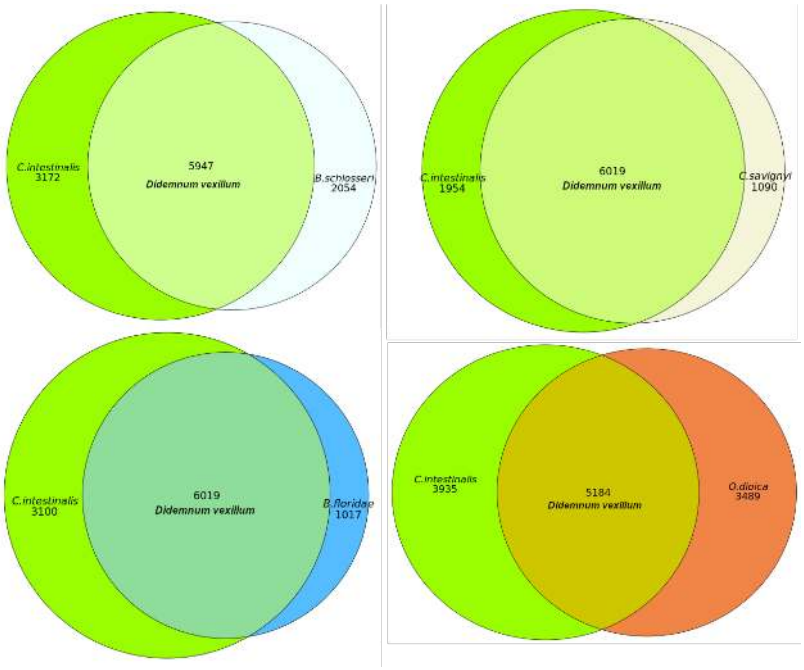


Figura 1-8.: Diagramas de Venn. La intersección representa las secuencias de proteínas con homología entre *D. vexillum* y *C. intestinalis*-*B. schlosseri*, *C. intestinalis*-*C. savignyi*, *C. intestinalis*-*O. dioica* *C. intestinalis* y el cefalodordado *B. floridae*

2. Identificación de subregiones sinténicas asociadas al sistema inmune innato del genoma de *Didemnum vexillum*

En este capítulo se presentan los resultados de la búsqueda de bloques sinténicos asociados a ontologías de genes putativos relacionados con el sistema inmune de la especie usando como referencia la anotación del fagosoma reportada en el KEGG. Para ello, en su orden, se describen los pasos y resultados para reportar por primera vez a la comunidad internacional, la primera versión de la anotación estructural y funcional para la especie.

Posterior al ensamblaje del transcriptoma descrito en el capítulo anterior, para el total de 90,938 transcritos ensamblados correspondientes a 39Mpb y con poder de codificación 17,540 regiones transcritas se procedió a la correspondiente anotación utilizando aproximaciones ab initio y datos de expresión propias del organismo. La pipeline Maker version 3.01.02 fue entrenada en una primera ronda utilizando los modelos existentes reportados para la especie modelo *Ciona intestinalis*. Este tunicado es considerado una especie modelo en muchos estudios de genómica y es la especie mejor estudiada no solo a nivel genómico sino la mejor anotada entre los genomas del grupo de los tunicados y por ende es generalmente utilizada como la especie de referencia para anotaciones. Adicionalmente para mejorar y validar los resultados intermedios obtenidos por la pipeline en el paso anterior, se incorporó una segunda ronda de entrenamiento utilizando la evidencia experimental del transcriptoma de la misma especie.

Para poder abordar la segunda parte de este capítulo, se menciona que aún con la importancia ecológica de los tunicados, los estudios genómicos son pocos y aún mas escasos los genómicos comparativos, en parte por la existencia de pocos genomas secuenciados por un lado y por otro por el estado de calidad de los genomas, los cuales principalmente se encuentran como versiones borradores. Hasta el momento se han anotado los genomas de tres ascidias solitarias: *C. savignyi*, *C. intestinalis* [Dehal *et al.*, 2002, Small *et al.*, 2007] y *O. dioica* [Denoeud *et al.*, 2010]; de los cuales el genoma de *C. intestinalis* fue mapeado en sus 14 cromosomas, en contraste a los organismos coloniales, donde solamente el genoma de *B. schlosseri* ha sido secuenciado, anotado y mapeado a sus 13 cromosomas [Voskoboynik *et al.*, 2013]. En los últimos años dos nuevos genomas han sido reportados en

la literatura los genomas borradores para las especies *Salpa thompsoni* y *Botrylloides leachii* [Jue *et al.*, 2016a], [Blanchoud *et al.*, 2018b] los cuales no fueron incorporados en este análisis pero análisis posteriores con estas especies permitirán completar los estudios comparativos del grupo.

Es importante mencionar que para realizar análisis sinténicos existen tres niveles del análisis genómico que incrementan la complejidad de los estudios genéticos comparativos: primero, el tiempo de divergencia entre las especies a comparar, segundo, las tasas evolutivas rápidas o lentas entre especies y tercero, los fenómenos de reorganización genómica posterior a la divergencia entre especies. Por ejemplo, recientes estudios en filogenómica comparativa de especies del subphylum [Delsuc *et al.*, 2018], indican que para los cuatro clados reportados (1) Appendicularia, (2) Thaliacea + Phlebobranchia + Aplousobranchia, (3) Molgulidae y (4) Styelidae + Pyuridae la diversificación data entre 450 y 350 millones de años atrás. Además, estudios enfocados en los análisis proteómicos previos para las especies *C. intestinalis* y *O. dioica* indican patrones y tasas de evolución peculiares, por ejemplo en Ciona se reporta tasas de evolución 50 % más altas cuando se comparan con otros vertebrados y aun sorprendente en Oikopleura para la cual se reporta tasas de evolución tres veces más rápidas de los observado para Ciona [Berná y Alvarez-Valin, 2014]. Estos trabajos fueron igualmente validados recientemente para los tunicados en las especies *S. thompsoni* y *B. leachii* [Jue *et al.*, 2016a],[Blanchoud *et al.*, 2018b]. Adicionalmente, los niveles de organización genómica son contrastantes en el grupo, ya que para el caso de *O. dioica* se presenta compactación de genoma debido no solo a la pérdida de elementos repetitivos sino debido a la reducción de las regiones intergénicas y a la reducción del tamaño de los intrones pero contrasta con cuyos niveles de reorganización en *C. intestinalis* que son reportados mas similares a los observados en los vertebrados [Berná y Alvarez-Valin, 2014].

Los análisis genómicos comparativos para identificar los bloques sinténicos se realizaron usando los algoritmos implementados en el programa Satsuma [Grabherr *et al.*, 2010]. Se escogió este software basado en el balance equilibrado que ofrece entre alta sensibilidad en la detección de bloques y la velocidad de ejecución. Es importante resaltar que aunque existen varias opciones de computo de bloques sinténicos [Haas *et al.*, 2004, Wang *et al.*, 2012, Proost *et al.*, 2012, Drillon *et al.*, 2014a]y, por otro lado las límitaciones y dificultades reportadas en los análisis sinténicos establecidos en la literatura [Liu *et al.*, 2018], el balance que describe el programa Satsuma permitió en un tiempo de ejecución razonable de maquina resolver las comparaciones pareadas de 109.769 contigs de *D. vexillum* con los miles de contigs de los genomas borrador de las demás especies de estudio.

Organismo	Tamaño estimado del genoma	Tamaño del Ensamblaje	Secuencias repetidas	N50	Asignación de Cromosoma	Método de secuenciación/-Tecnología	Método de Ensamblaje Genómico	Citas
B. schlosseri	725Mbp	580Mbp	Integrado	38kbp (longitud del contig del cromosoma)	40 % de los genes fueron asignados a distintos cromosomas	Colonia Sc6a-b - >Protocolos single read y paired ends secuenciados bajo las plataformas 454GS-FLX e Illumina Genome Analyzer II (GA II). Tamaño del inserto 400-600bp.	Newbler v2,5 (Rche) parámetros por defecto, modo heterocigoto.	[Voskoboynik *et al.*, 2013]
C. Intestinalis	160Mbp	116.7Mbp	Removido	n/a	La asignación de genes fue obtenida varios años seguido de la publicación inicial	librerías BAC y Cosmid, método shotgun, secuenciamiento por Sanger	Suit de ensamblaje JAZZ	[Dehal *et al.*, 2002]
C. Savignyi	180Mbp	590Mbp	n/a	601kbp (scaffold)	n/a	método shotgun, secuenciamiento por Sanger	-	[Small *et al.*, 2007]
O. dioica	65Mbp	70.5Mbp	n/a	395kbp (scaffold)	n/a	método shotgun, secuenciamiento por Sanger	-	[Seo *et al.*, 2001]
Branchiostoma floridae	500Mpb	-	Removidos	scaffold N/L50 = 62/2.6 Mb and contig N/L50 = 4916/28kb	n/a	Librerías BAC (CHORI-302), método shotgun, secuenciamiento por Sanger	Suit de ensamblaje JAZZ	[Putnam *et al.*, 2008]

Tabla 2-1.: Genomas de los organismos ensamblados.

2.1. Hacia una Anotación Funcional y Estructural

Los datos generados del genoma de *D. vexillum*, su mRNA y los transcritos no son información directamente usable sin una de anotación genómica que genere resultados que permitan medir la relevancia de las caracteristicas biológicas; de otro modo, el número de estructuras de proteínas y secuencias genómicas no caracterizadas aumentaría [Reeves *et al.*, 2008]. En algunos casos la anotacion es una tarea especializada de bases de datos tales como Ensembl [Birney *et al.*, 2004], limitado a genomas de vertebrados, VectorBase [Lawson *et al.*, 2006] restringido a insectos transmisores de enfermedades humanas; para los tunicados, **ANISEED** [Brozovic *et al.*, 2018] [1] es la base de datos que da acceso a los genomas anotados de catorce especies [aniseed, 2020] con los que usando genómica comparativa y las secuencias anotadas estructural y funcionalmente de *D. vexillum* nos puede conducir a elucidar el papel que juegan los genes que componen el genoma [Eilbeck *et al.*, 2005].

2.1.1. Anotación Genómica Estructural

Debido a la cantidad de secuencias a examinar y el volumen de datos que se pueden generar, la *pipeline* MAKER[Cantarel *et al.*, 2008] fue usada para anotación genómica estructural del ensamble genómico de *D. vexillum*. Este software es un flujo computacional validado con el genóma del organismo *Caenorhabditis elegans*[2] y se uso en la anotación estructural en la base de datos del genoma de *Schmidtea mediterranea*[3]. MAKER automatizó la anotación con indices de calidad que usan evidencia experimental para postular modelos candidatos a genes en donde identifica elementos como intrón, exon, CDS, regiones repetitivas, UTR entre otros. Maker requiere como datos de entrada para cada gen a modelar; una **evidencia de proteína**, **una evidencia de transcrito** y **una evidencia de genoma**.
La Anotación estructural con Maker tiene cinco fases por las que pasan todas las secuencias candidatas a ser caracterizadas.

1. **Fase de cómputo**

   El análisis de cada entrada genómica comienza con identificar y localizar los elementos repetitivos de baja complejidad y secuencias repetitivas intercaladas [Abouelhoda *et al.*, 2002]: las secuencias identificadas como de baja complejidad son enmascaradas en la entrada genómica sin afectar el marco de lectura o el tamaño en la secuencias de DNA pero, son excluidas de posteriores alineamientos de BLAST con la evidencia de proteína y de transcrito. En esta etapa la identificación de secuencias repetitivas *de novo* se hizo en un procedimiento independiente(Anexo G) a Maker y usando el software RepeatModeler v.1.0.4 [Smit y Hubley, 2019] con los parámetros por omisión y usando los

[1] `https://www.aniseed.cnrs.fr`
[2] `http://www.wormbase.org/`
[3] `http://smedgd.neuro.utah.edu`

scaffolds del ensamble genómico de *D. vexillum*. El enmascaramiento de las repeticiones y las secuencias de baja complejidad se realizó con RepeatMasker v.open-4.0.5 [Smit *et al.*, 2015] usando la libreria de repeticiones de novo en combinación con la librerias de repeticiones de *C. intestinalis* de RepBase Version 20.03 [Bao *et al.*, 2015]. La libreria de elementos repetitivos conocidos(anotados) fue un parámetro de entrada en Maker.

Alineamiento de las secuencias DNAg, mRNA y Proteína de *D. vexillum*: hasta aquí el tratamiento dado a las secuencias genomicas busca identificar regiones codificantes de proteínas y, dado que la predicción *in sílico* no es completamente confiable debido a la poca evidencia experimental que existe a la fecha del anteproyecto que valide los resultados, en esta etapa y en las subsecuentes MAKER toma ventaja de los alineamientos entre las secuencia del DNAg, el mRNA y la proteína para validar cada modelo computacional, un alineamiento significativo entre los tres elementos es uno de lo indices de calidad para cada modelo de gen.

Los algoritmos como BLAST[Altschul *et al.*, 1990] son una buena solución para la comparación de largas secuencias genómicas, pero realizar una búsqueda exhaustiva de homología con él implica recursos de ejecución (tiempo computacional); de modo que esta no es la solución heurística óptima para buscar pequeñas secuencias conservadas, tales como sitios donores y aceptores o sitios de empalme en los límites intron-exon en largas secuencias: Blast no toma en cuenta los sitios de *splicing*, por lo que sus alineamientos son una aproximación en bruto para que MAKER use el *software* de alineamientos Exonerate[Slater y Birney, 2005], que incluye el algoritmo "protein2genome" que integra la predicción de sitios de splicing permitiendo alineamientos con el modelamiento de las regiones intrónicas.

2. Filtrado y Agrupación

   En esta fase, Maker identifica y remueve predicciones marginales o secuencias con alineamientos sobre puntajes básicos p. eje. bajos porcentajes de identidad. Los criterios para el filtrado son parte del archivo maker_bopts.ctl. Los alineamientos no excluidos son solapados con las secuencia genómica de *D. vexillum* con dos propósitos: agrupar los alineamientos y los modelos de genes que tiene diferentes orígenes (snap, genemark, augustus) para validarlos e identificar evidencia redundante.

3. Pulido

   Este paso realiza un re-alineamiento de los encuentros obtenidos con BLAST con el algoritmo Exonerate, para obtener mayor precisión en los límites de los exones; generando informacion más precisa sobre los sitios donores y aceptores de splicing que será útil en la siguiente fase de MAKER.

4. Síntesis

Maker sintetiza la información del paso anterior, agrupando los alineamientos de mRNA y proteína por cada modelo de gen, para reportar evidencia de las anotaciones. Para hacer esto, identifica los mRNA que se originan por splicing alternativo desde las secuencias de un mismo gen.

En esta fase Maker identifica los nucleótidos de la secuencia genómica que corresponden a las secuencias intergénicas. Basado en los alineamientos de proteína y mRNA se identifica su sentido/antisentido codificantes etiquetándolos como posibles regiones intergénicas.

MAKER luego calcula un score sobre la secuencias que dan alineamiento(secuencias codificantes) basado en el porcentaje de similaridad del alineamiento, el tipo de alineamiento y la posicion de los nucleótidos en el alineamiento. El score junto a sus secuencias putativas(regiones intergénicas, regiones codificantes, intron y UTR's) son pasadas a SNAP. Basado en esta información , SNAP modifica su perfil de Modelo Oculto de Markov(HMM) para la predicción del gen de forma más certera.

5. Anotación

   Maker post-procesa las predicciones de SNAP generadas en la fase de síntesis y las combina con las evidencias para completar la anotación. Cada predicción de SNAP es chequeada nuevamente para que los mRNA's y las secuencias UTR 5' y 3' sean consistentes con la predicción e identificación de los exones codificantes, de esta forma las coordenadas de las predicciones de SNAP son alteradas para incluir estas regiones. Este proceso es repetitivo para cada modelo proveniente de la fase de síntesis para, finalmente procesar la evidencia documentando cada uno de los exones del modelo.

Maker concluye todo el trabajo de anotación estructural y entrega los resultados en un formato plano de texto, delimitado por tabulaciones y conocido como GFF3. Los resultados obtenidos se puede visualizar en la figura **2-1**.

Numero de Genes	62194	Total gene length	108000074
Number of mrnas	64401	Total mrna length	112120628
Number of mrnas with utr both sides	12553	Total cds length	30927573
Number of mrnas with at least one utr	28036	Total exon length	35972707
Number of cdss	64401	Total five_prime_utr length	1809104
Number of exons	169755	Total three_prime_utr length	3236030
Number of five_prime_utrs	25043	Total intron length per cds	73640933
Number of three_prime_utrs	15546	Total intron length per exon	76253275
Number of exon in cds	165544	Total intron length per five_prime_utr	1322530
Number of exon in five_prime_utr	26964	Total intron length per three_prime_utr	1236952
Number of exon in three_prime_utr	17748	Longest cds piece	9852
Number of intron in cds	101143	Longest five_prime_utr piece	4214
Number of intron in exon	105354	Longest three_prime_utr piece	4569
Number of intron in five_prime_utr	1921	Longest intron into cds part	18681
Number of intron in three_prime_utr	2202	Longest intron into exon part	18681
Number of single exon gene	17261	Longest intron into five_prime_utr part	17518
Number of single exon mrna	18378	Longest intron into three_prime_utr part	8943
mean mrnas per gene	1	Shortest genes	12
mean cdss per mrna	1	Shortest mrnas	12
mean exons per mrna	2.6	Shortest cdss	3
mean five_prime_utrs per mrna	0.4	Shortest exons	1
mean three_prime_utrs per mrna	0.2	Shortest five_prime_utrs	1
mean exons per cds	2.6	Shortest three_prime_utrs	1
mean exons per five_prime_utr	1.1	Shortest cds piece	1
mean exons per three_prime_utr	1.1	Shortest five_prime_utr piece	1
mean introns in cdss per mrna	1.6	Shortest three_prime_utr piece	1
mean introns in exons per mrna	1.6	Shortest intron into cds part	5
mean introns in five_prime_utrs per mrna	0	Shortest intron into exon part	5
mean introns in three_prime_utrs per mrna	0	Shortest intron into five_prime_utr part	14
mean gene length	1736	Shortest intron into three_prime_utr part	7
mean mrna length	1740	mean five_prime_utr length	72
mean cds length	480	mean three_prime_utr length	208
mean exon length	211	mean cds piece length	186

Figura 2-1.: Resultados de la anotación estructural con la pipeline MAKER y secuencias de *D.vexillum*

2.1.2. Uso de Ontología para anotación estructural

Para evitar ambigüedades en la terminología biológica Maker usa SOFA(Sequence Ontology Feature Annotation) [Eilbeck *et al.*, 2005], un subconjunto de términos y relaciones del GO [Consortium, 2006] que pueden ser anotados a las secuencias genómicas identificadas. Los resultados para todos los candidatos a genes se visualizan en la base de datos como en la figura **2-2**, donde el navegador genómico Gbrowse visualiza la anotación estructural y algunas ontologias SOFA asociadas al candidato a gen, gene,exon y repetición y cuyo proceso de implementación será presentado en el siguiente capítulo.

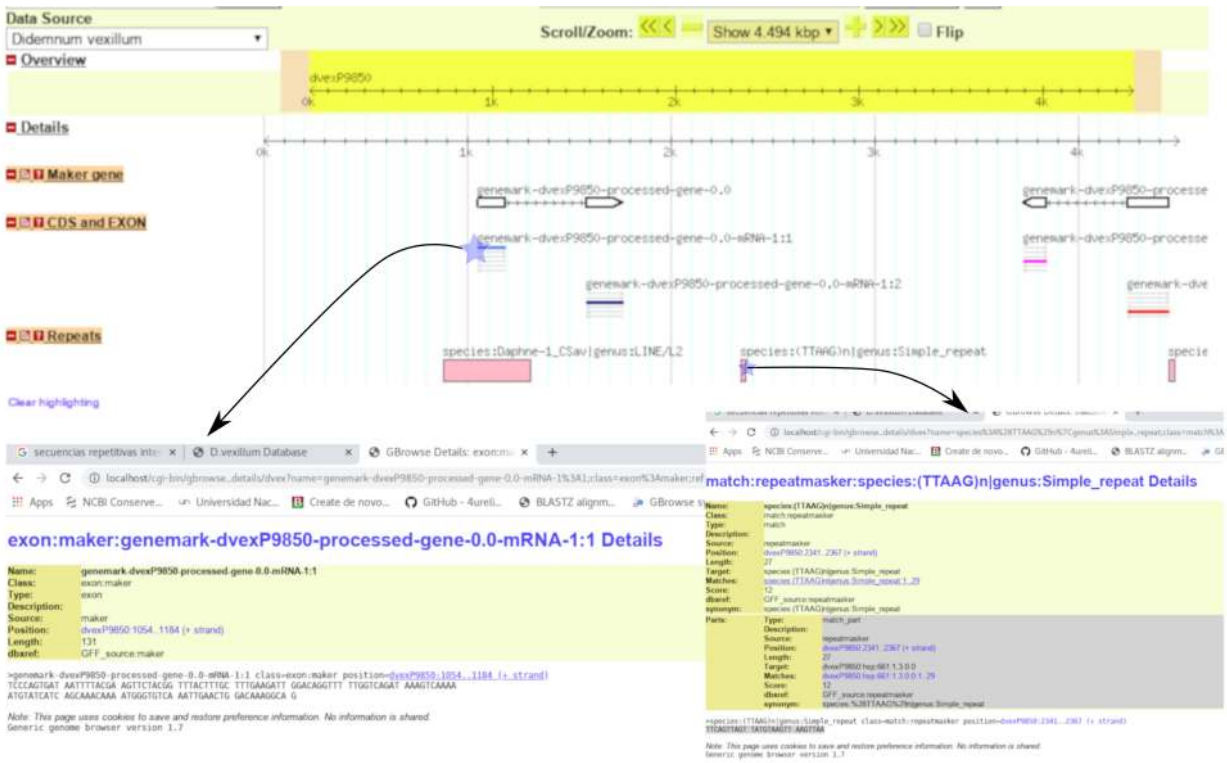


Figura 2-2.: Captura de pantalla de la Base datos *D.vexillum* mostrando resultados de Maker.

2.1.3. Anotación Genómica Funcional

La identificación de elementos funcionales en las secuencias candidatas a proteinas de *D. vexillum* generadas con Maker, se realizó con el software Interproscan [Quevillon *et al.*, 2005], usando la base Uniprot90 con un valor e-value de 1 E-25. Los resultados para la anotación se resumen en la figura **2-3** que muestra el numero de registros de la base de datos de Uniref90 en Uniprot[4] desde los que se obtuvo homología y se transfirió el mayor porcentaje

[4] `https://www.uniprot.org`

de la anotación con proteínas que tienen por lo menos un porcentaje de identidad del 90 % del género *Ciona*, un porcentaje menor fue anotado por homología de mínimo un 90 % de identidad con proteínas originadas en la familia *Branchiostomidae*.

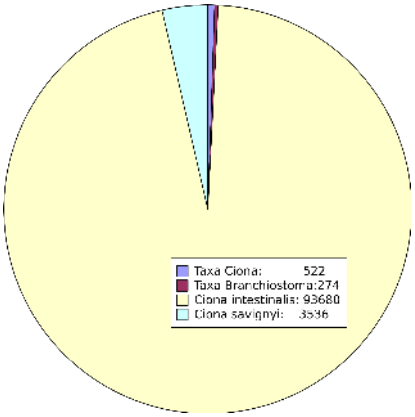


Figura 2-3.: Resultados de la anotación funcional con la base de datos Uniref90 de Uniprot con las secuencias de *D.vexillum*

Con el fin dar soporte a la anotación de genes asociados al sistema inmune de *D. vexillum*, de los resultados de la anotación funcional con Interproscan(PFAM) se clasificaron las ontologías transferidas por homología entre proteínas y luego se filtraron por *Clase Inmune* y *Clase Base* de acuerdo a la descripción hecha en [Hu *et al.*, 2008]. La figura **2-4** visualiza los resultados.

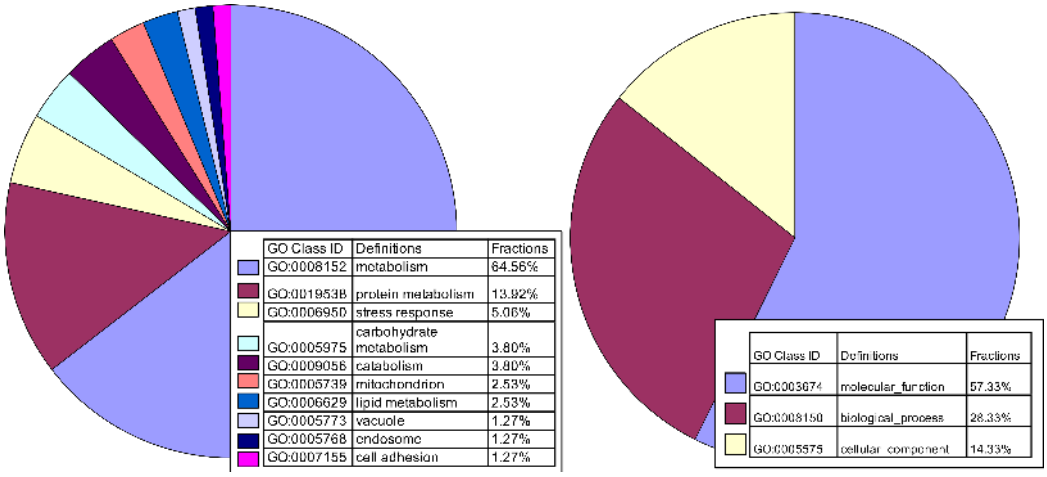


Figura 2-4.: Clasificación por familias de ontologías anotadas a 4371 secuencias de proteínas de *D.vexillum*(evalue 1E-24) con Interproscan.

2.2. Microsintenias asociadas al Sistema Inmune Innato

La Genómica Comparativa menciona la sintenia cuando hace referencia a la organización genómica conservada, en donde el *loci* de ortólogos genéticos esta físicamente co-localizado en el mismo scaffold o cromosoma en dos o mas especies: esta similaridad se debe a que se comparten las caracterizaciones genómicas con un ancestro común.

De acuerdo a una escala de "conservación sinténica" esta puede ser clasificada en **Macrosintenia** o **Microsintenia**. La Macrosintenia es el término que hace referencia a una conservación mayoritariamente a nivel cromosomal; donde el *loci* compartido esta el mismo cromosoma, aun cuando la localización y el orden pueda variar entre especies. La Microsintenia de otra parte, señala la organización genómica a una escala más pequeña, en este caso, el orden conservado en los genes vecinos es frecuentemente el mismo, comúnmente llamada *colinearidad*[Zhao y Schranz, 2019].

Con el objetivo de identificar subregiones del sistema inmmune asociadas a regiones sinténicas compartidas entre tunicados, se usaron los genes asociados con la ruta metabolica del fagosoma del *C. intestinalis*: cin04145 anotado en el KEGG y se usó como referencia debido a que es el único camino biológico asociado al sistema inmune innato. Además porque la fagocitosis es un proceso altamente conservado que surgió antes del desarrollo de la multicelularidad [Stuart y Ezekowitz, 2008] y representa un evento temprano mediado por el sistema inmune innato y crucial que activa las defensas del huésped contra los patógenos invasores que inicia con la unión y el reconocimiento de partículas por los receptores de la superficie celular [Henneke y Golenbock, 2004]. De hecho, los fagocitos en la túnica en los Tunicados, se encuentran en todas las ascidias conocidas, por lo que, a los fagocitos se les asocia la hipótesis de ser basales y compartidos con tunicados ancestrales[Hirose, 2009]

La identificación de homología entre las proteínas de *D. vexillum* obtenidas en la anotación estructural con Maker y ochenta cadenas proteicas que constituyen el fagosoma de *C. intestinalis* sirvieron a un primer acercamiento usando el software Blast con un valor de e-value de 1e-25 y seleccionando el primer acierto: se obtuvo homología con 69 proteinas(86 %) de un cubrimiento promedio del 70 % y cuyos resultados se visualizan en la figura **2-5**.

Los bloques sinténicos entre tunicados se identificaron con el software Satsuma [Grabherr, 2010], y su análisis se organizó en tres niveles comparativos para identificar los bloques sintenicos compartidos homólogos al sistema inmune innato(fagosoma); un primer bloque de sintenia conservada a nivel pareado especie *D. vexillum* y *C. intestinalis*, en el segundo nivel se encuentran los bloques de sintenia conservado entre *D. vexillum* y las especies *C. intestinalis*, *O. dioica*, *C. savignyi* y *B. schlosseri*. El tercer y ultimo nivel compara los bloques comunes a los tunicados del nivel dos y el cefalocordado *B. floridae*.

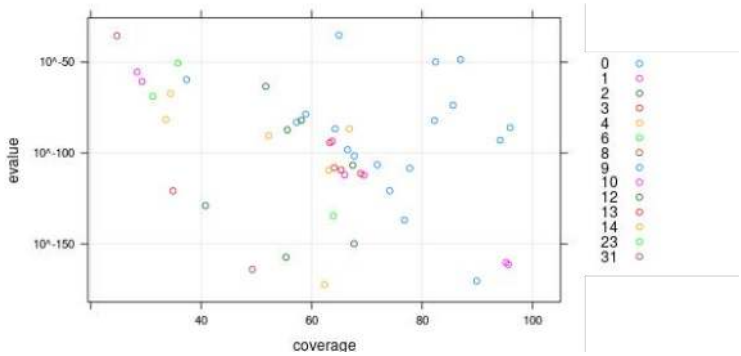


Figura 2-5.: Homologia entre 69 proteinas de D.vexillum y 80 proteinas del fagosoma de C. intestinalis, e-value=1e-25.

2.2.1. Bloques de sintenia a nivel de DNAg entre *D. vexillum* y *C. intestinalis*

512 Bloques de sintenia fueron identificados entre todo el genoma de *C. intestinalis* y todo el genoma de *D. vexillum*, con un promedio de cubrimiento del 58 % distribuidos y ubicados sobre las coordenadas de cada cromosoma de *C. intestinalis* que se describen en las figuras **2-6**, **2-7** y **2-8**.

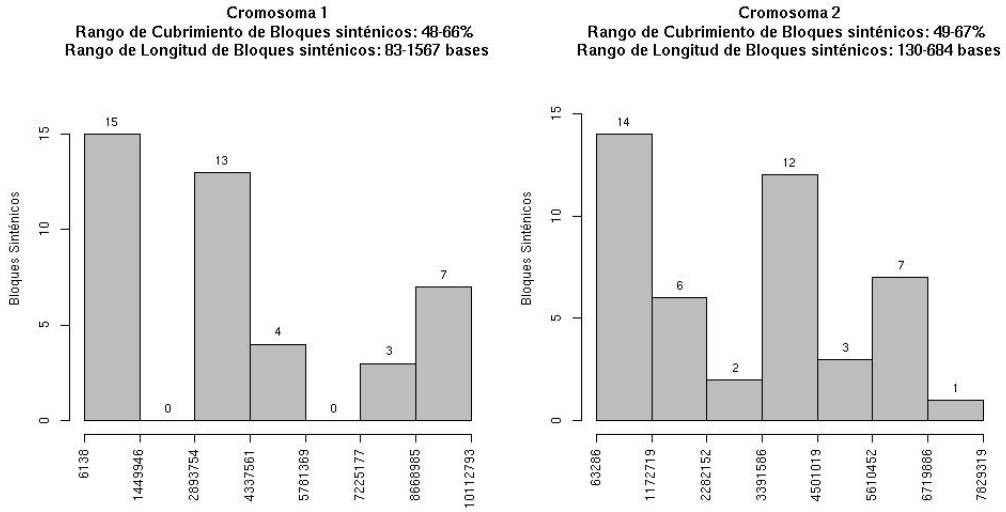


Figura 2-6.: Bloques sintenicos entre *D.vexillum* y *C. intestinalis*, e-value=1e-25.

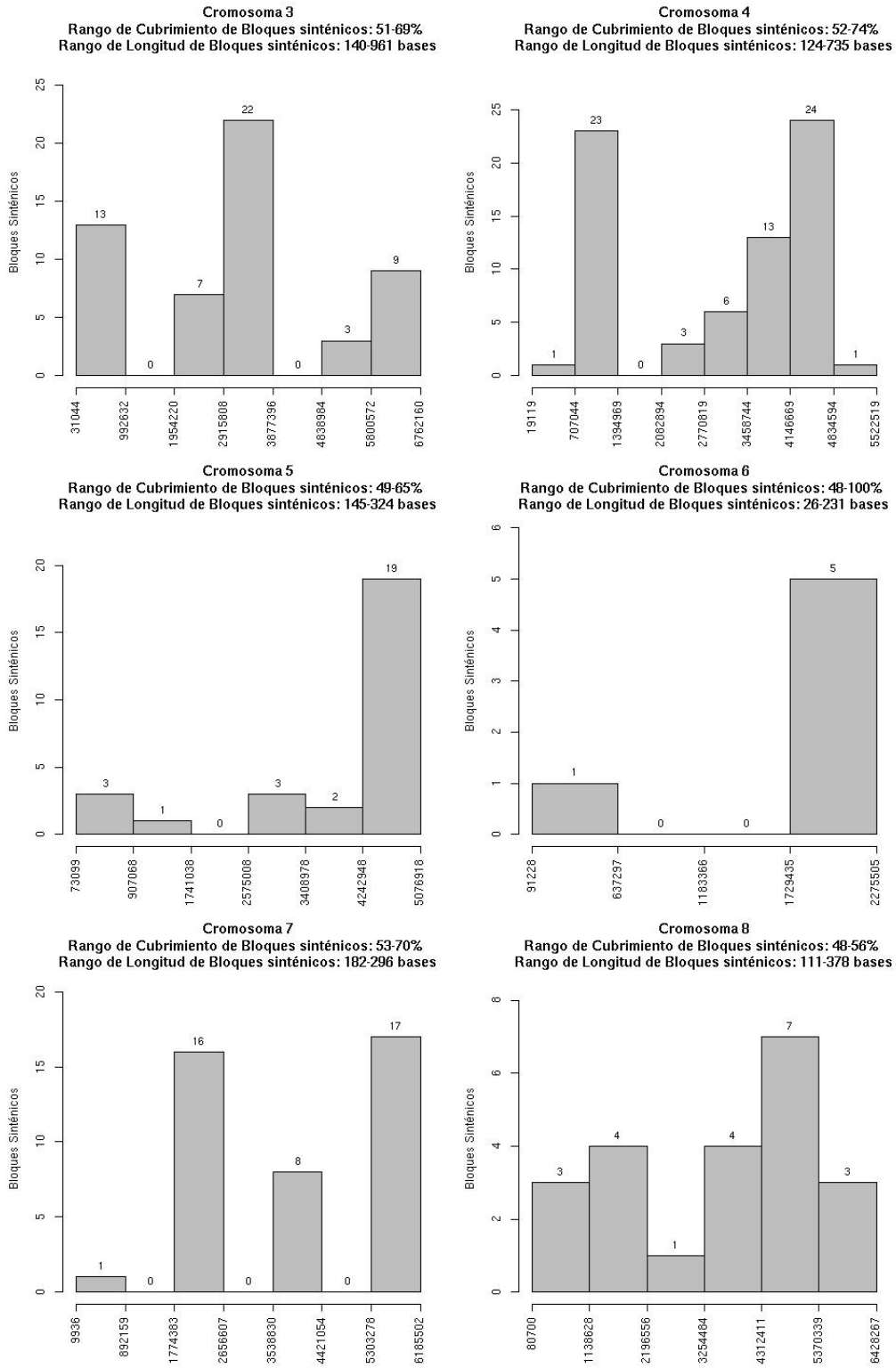


Figura 2-7.: Bloques sintenicos entre *D.vexillum* y *C. intestinalis*, e-value=1e-25.

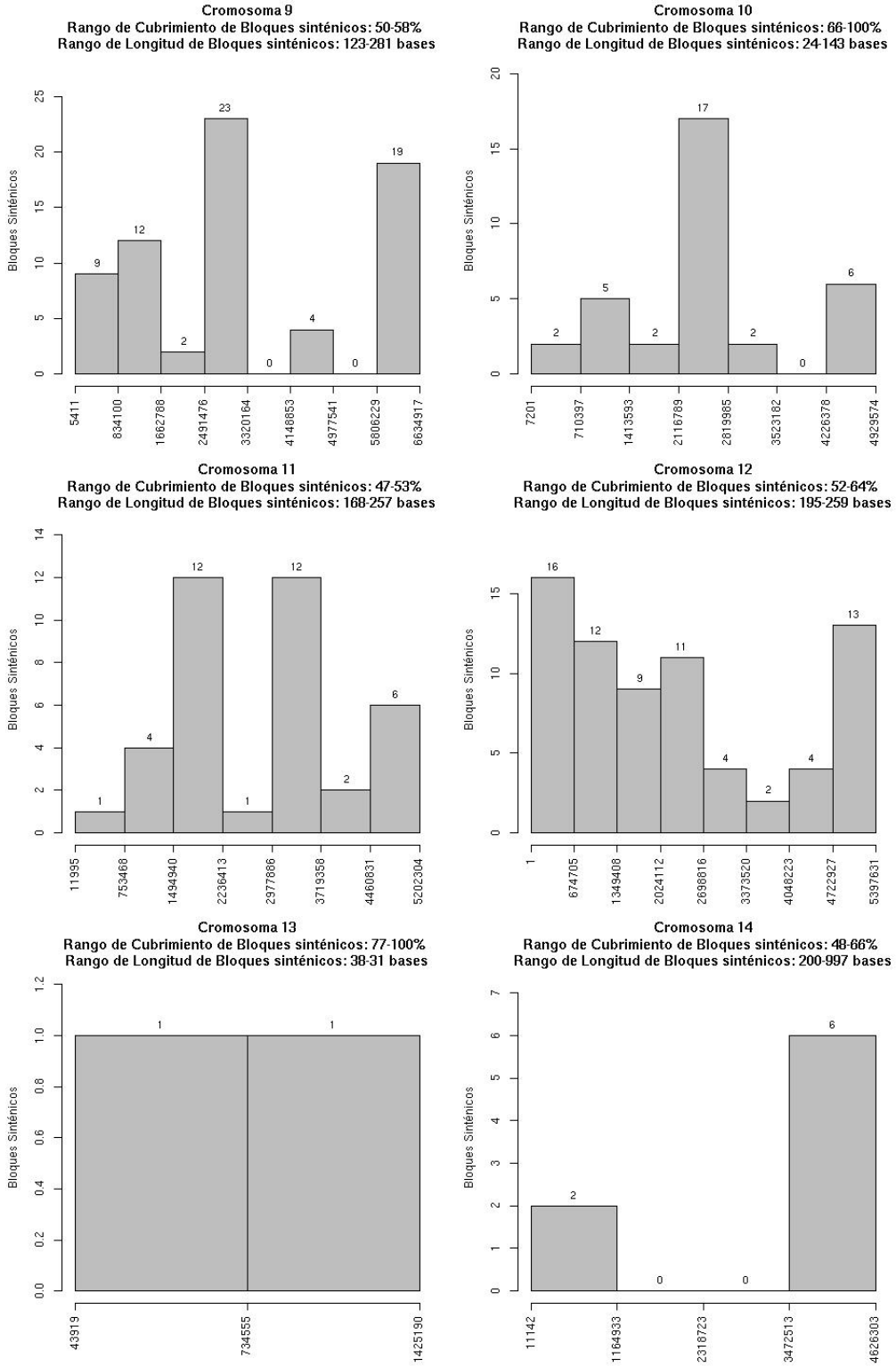


Figura 2-8.: Bloques sintenicos entre *D.vexillum* y *C. intestinalis*, e-value=1e-25.

2.2.2. Bloques de sintenia conservada a nivel de DNAg entre *D. vexillum* con *C. savignyi*, *B. schlosseri*, *O. dioica* y el cefalocordado *B. floridae*

Los datos de microsintenia entre *D. vexillum* y los catorce cromosomas de *C. instestinalis* fueron preparados para ser visualizados dinámicamente(Figura **2-9**) con la herramiento Mizbee [Meyer *et al.*, 2009]. Este mismo procedimiento se realizó con las comparaciones entre *D. vexillum* con *C. savignyi*(Figura**2-10**),*B. schlosseri*(Figura **2-11**, *O. dioica*(Figura **2-12**), y el cefalocordado *B. floridae*(Figura **2-13**). En cada figura se observa en el anillo extremo el genoma usado como fuente, es decir la información genómica organizada en cromosomas, raftigs, scaffolds o contigs según el tipo de información genómica disponible de cada especie *C. intestinalis*, *C. savignyi*, *B. schlosseri*, *O. dioica* y *B. floridae* respectivamente. En el anillo interno se ubican los scaffolds de destino de la especie *D. vexillum* y el mejor fragmento con el cual se comparte la sintenia de la especie fuente. Las conexiones con mayor conservación son codificadas con colores y la lineas puntualmente representan conexiones específicas para relaciones de genes del sistema inmune entre los genomas comparados, en particular con el fragmento genómico con mejores relaciones sinténicas. Por otro lado, en la primera imagen de la derecha, se observa con mejor detalle la relación capturada en el anillo interno, en la cual los bloques de sintenia compartida se observan como bloques de colores y en la segunda imagen un detalle de orden de genes compartidos a mejor resolución. Dependiendo del estado del genoma se puede observar con mayor claridad la relaciones para *D. vexillum* y las especies *C. intestinalis* y *B. schlosseri*.

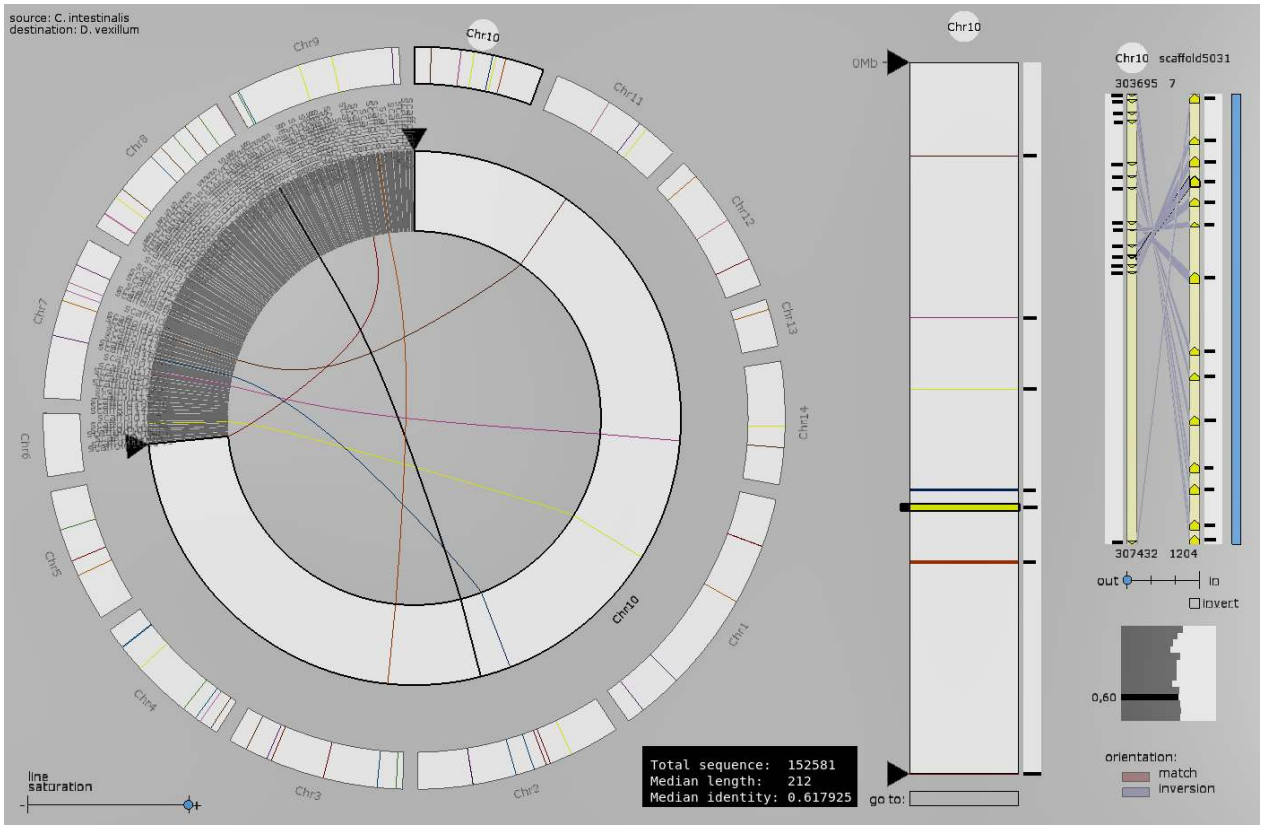


Figura 2-9.: Microsintenia entre *D.vexillum* y *C. intestinalis*

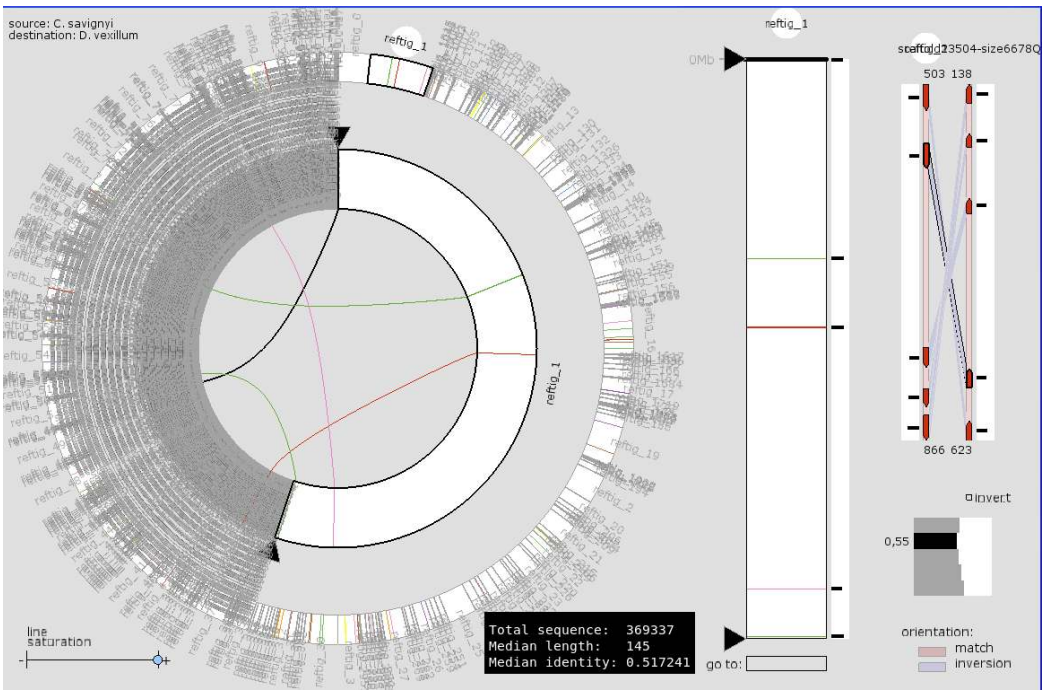


Figura 2-10.: Microsintenia entre *D.vexillum* y *C. savignyi*

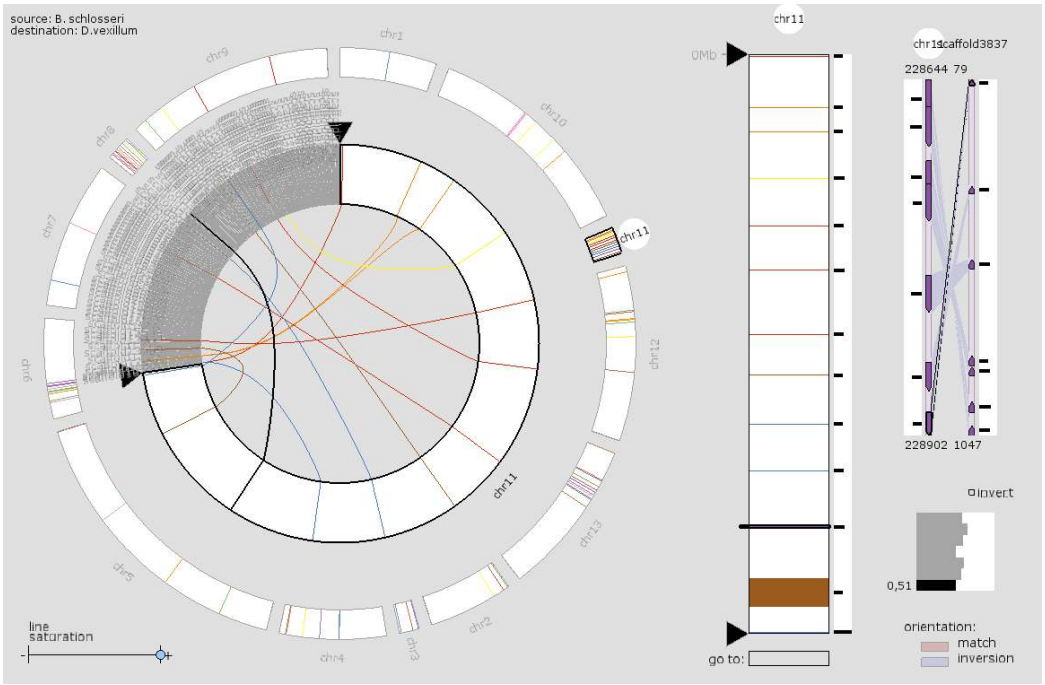


Figura 2-11.: Microsintenia entre *D.vexillum* y *B. schlosseri*

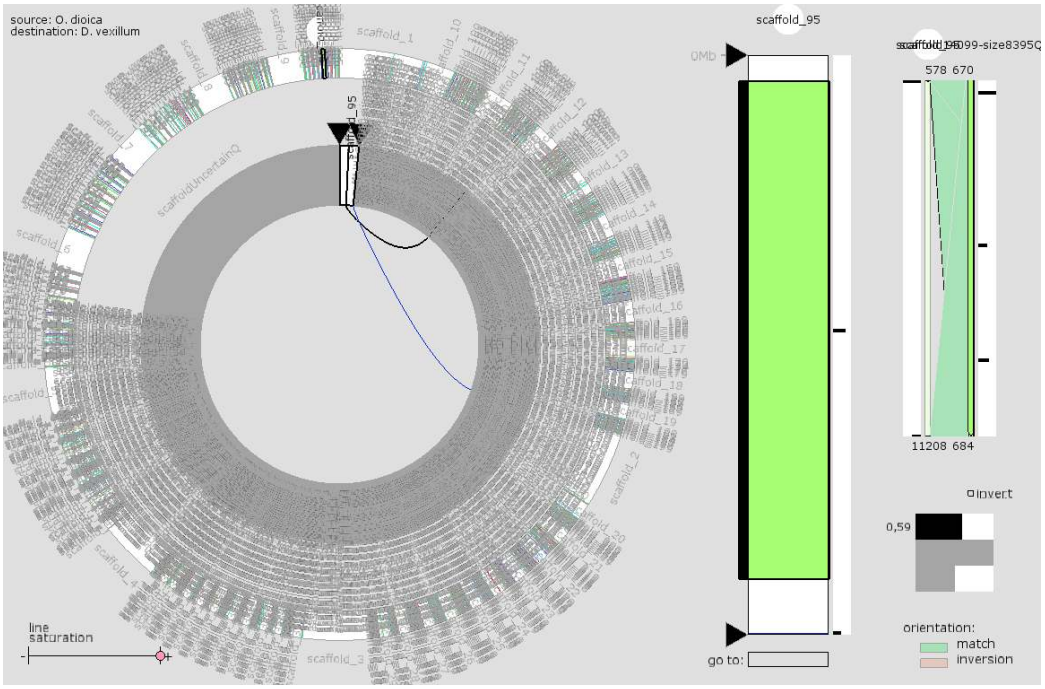


Figura 2-12.: Microsintenia entre *D.vexillum* y *O. dioica*

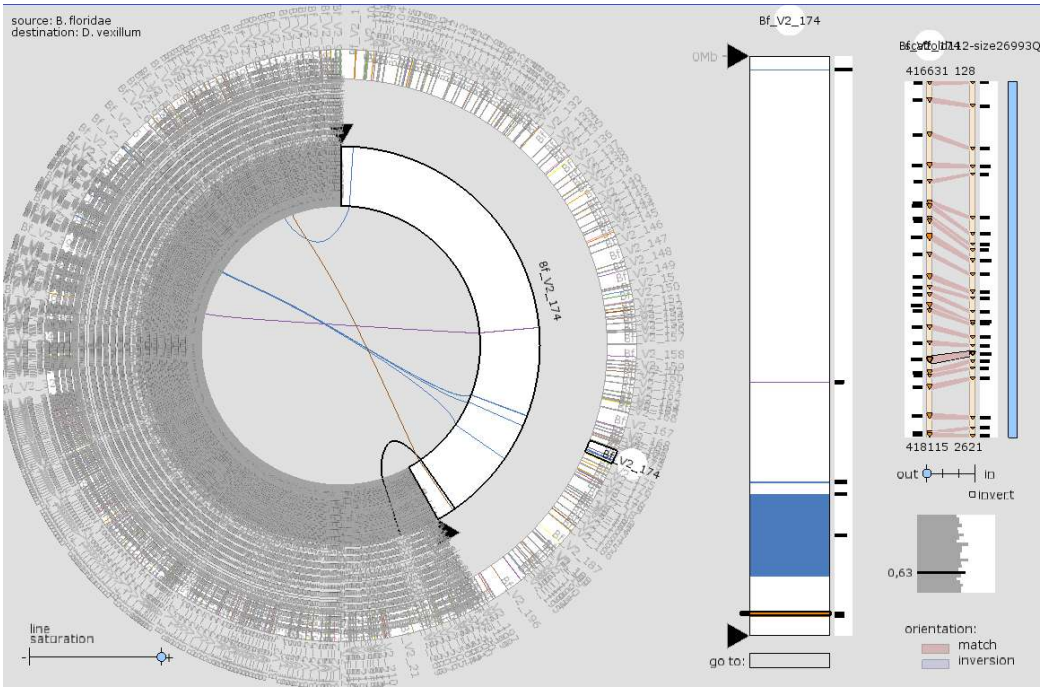


Figura 2-13.: Microsintenia entre *D.vexillum* y *Branchiostoma floridae*

2.2.3. Identificación de la colinearidad entre las secuencias genómicas de *D. vexillum* y otros tunicados

A nivel de DNAg nuestros organismos de interés divergen, sin embargo a nivel de proteínas su ubicación en el genoma, es decir, la colinearidad, junto con la identificación de los genes que evolutivamente son mas conservados (ortólogos) y la permisividad de excluir sus regiones intergenicas (regiones menos conservadas), nos permite encontrar fragmentos de genes candidatos a codificar proteínas (incluyendo las del fagosoma) entre nuestras especies.

Identificación de genes en microsintenia.

En un primer paso se obtuvieron los candidatos a genes ortólogos con el software OPSCAN que usó el método de los mejores aciertos recíprocos(RBH)[5] entre las secuencias proteicas de los organismos *D. vexillum*, *C. intestinalis(GCA_000224145.2)*, *C. savigyni(Emsembl, v99)*, *O. dioica(NCBI ASM20955v1)*, *B. leachi, B. schloserii(unicamente 68 % del proteoma)* y *B floridae(GCF_000003815.1 V2)*, la herramienta OPSCAN usada está incluida en el software SynChro[Drillon *et al.*, 2014b]. Los resultados se resumen en la tabla 2.2.3 y fueron la entrada al software i-adhore [Simillion *et al.*, 2008] que se usó para identificar la colinearidad en el orden de los genes.

	D. vexillum	*C. intestinalis*	*C. savignyi*	*O.dioica*	*B. schlosseri*	*B. floridae*
Proteina Total	12561	21096	20155	13505	46519	28623
D. vexillum	–	288	275	35	97	79
C.intestinalis	288	–	2617	66	197	219
C. savignyi	275	2617	–	55	213	184
B. schlosseri	97	197	213	28	–	88
B. floridae	79	219	184	50	88	–
O. dioica	35	66	55	—	28	50

Tabla 2-2.: Mejor Blast Recíproco con similaridad mayor a 82 % y una relación de longitud del 44 % para 5 especies de tunicados y B. floridae.

Identificación de Microsintenia.

Un cruce de información entre: el orden conservado de las secuencias de proteínas(I-adhore) originado en los alineamientos con mejores aciertos recíprocos(Opscan) y, los datos generados del los alineamientos de todo el genoma(Satsuma) entre *D. vexillum* con *C. intestinalis*, *C. savigyni*, *O. dioica*, *B. schlosseri* y el cefalocordado *Branchiostoma floridae*, nos permitió inferir las secuencias genómicas conservadas entre los tunicados de estudio.

[5]Reciprocal Best Hits(BDBH)

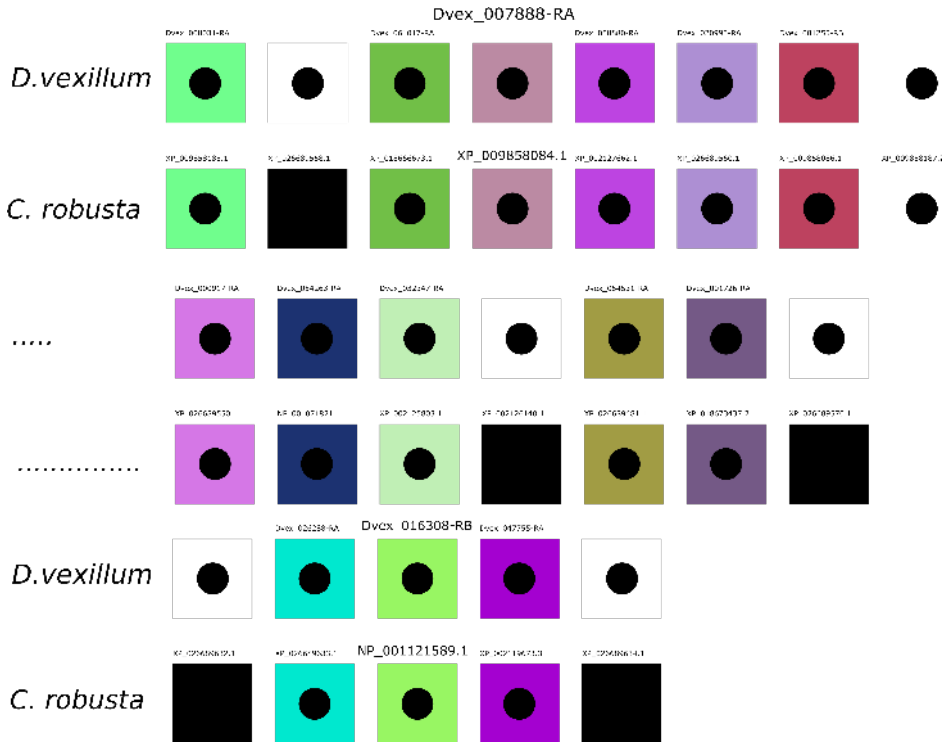
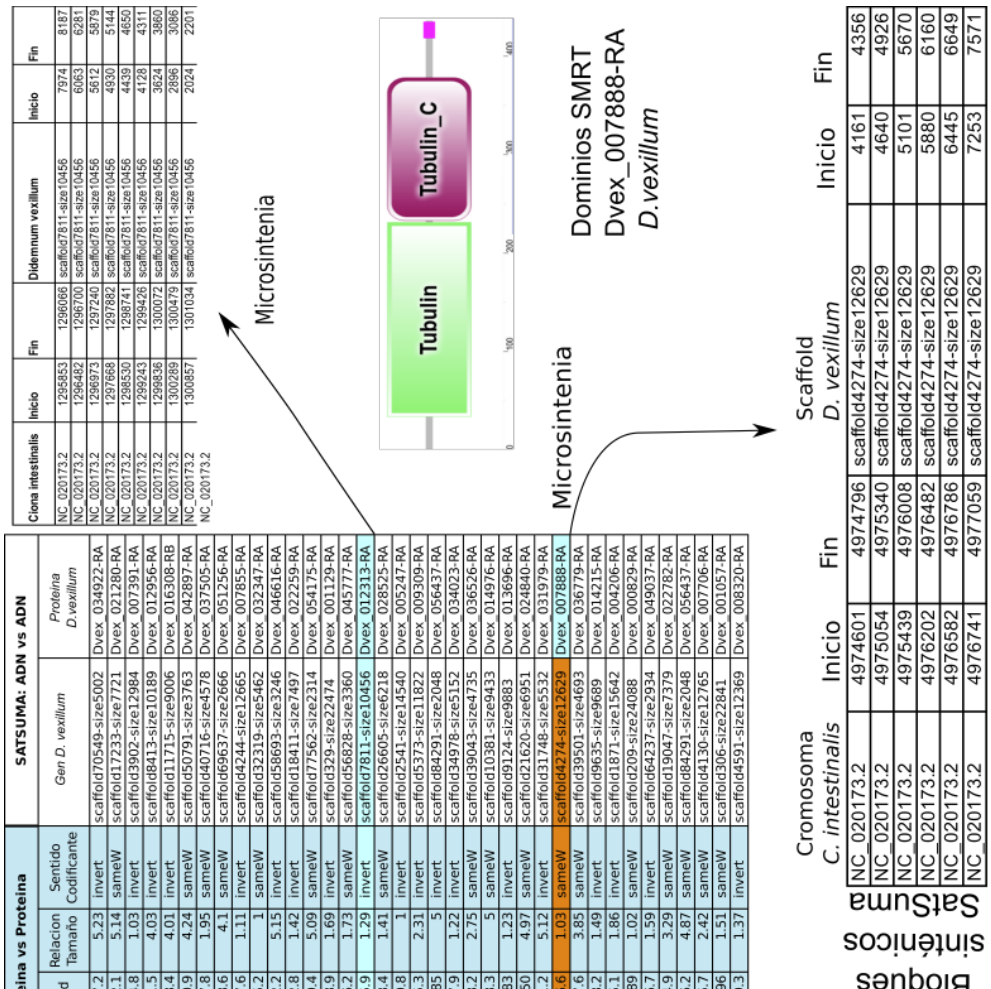


Figura 2-14.: Colinearidad de Peptidos de *D.vexillum* con referencia de *Ciona intestinalis*

Una descripción como ejemplo de los resultados se resume en la figura **2-17** y **2-17** para las especies *D. vexillum* y *C. intestinalis*, donde la proteina del fagosoma de *C. intestinalis* XP_009858084.1 muestra una similaridad por Blast recíproco de 96.6 % y tiene un 3 % de mayor tamaño con respecto a la proteína de *D. vexillum* Dvex_007888-RA, originada en el scaffold "scaffold4274size12629". La figura muestra una conservación de dominios entre las dos proteínas y la microsintenia genomica (Bloques sinténicos Satsuma) conservada con coordenadas en *C. intestinalis*, con respecto al scaffolds de *D. vexillum*.
Otras regiones con microsintenia son las que muestran las secuencias XP_009858084.1(tubulin alpha chain, testis-specific-like [C. intestinalis]) y XP_002129440.1 (tubulin alpha-1A chain [C. intestinalis]). Estos microtúbulos realizan diversas funciones que son esenciales en eucariotes: están compuestos por un heterodímero alfa y beta tubulina. Los genes que codifican estos constituyentes de microtúbulos son parte de la superfamilia de tubulina transversal a todos los eucariotes; aunque la similaridad en aminoácidos entre ortólogos de organismos lejanos evolutivamente se ha reportado entre 35-40 % y, aunque su similaridad con cualquier otra proteína es mínima [Little y Seehaus, 1988]. Las tubulinas son genes que expresan variantes de transcripción(*D.vexillum*, scaffold4274-size12629, codifica 6 proteinas) que codifican diferentes isoformas para estos genes.



I-adHORE Colinearidad Secuencias de Proteina		OPSCAN: Mejor Blast recíproco, Proteina vs Proteina				SATSUMA: ADN vs ADN		
D. vexillum	*C. intestinalis*	*D. vexillum*	Identidad	Relacion Tamaño	Sentido Codificante	*Gen D. vexillum*	*Proteina D.vexillum*	
Dvex_034922-RA	NP_001027676.1	Dvex_034922-RA	57.2	5.23	invert	scaffold70549-size5002	Dvex_034922-RA	
Dvex_021280-RA	NP_001027684.1	Dvex_021280-RA	62.1	5.14	sameW	scaffold17233-size7721	Dvex_021280-RA	
Dvex_007391-RA	NP_001028999.1	Dvex_007391-RA	54.8	1.03	invert	scaffold3902-size12984	Dvex_007391-RA	
Dvex_012956-RA	NP_001029015.1	Dvex_012956-RA	81.5	4.03	invert	scaffold8413-size10189	Dvex_012956-RA	
Dvex_016308-RB	NP_001121589.1	Dvex_016308-RB	68.4	4.01	invert	scaffold11715-size9006	Dvex_016308-RB	
Dvex_042897-RA	NP_001121595.1	Dvex_042897-RA	70.9	4.24	sameW	scaffold50791-size3763	Dvex_042897-RA	
Dvex_037505-RA	XP_002121757.1	Dvex_037505-RA	77.8	1.95	sameW	scaffold40716-size4578	Dvex_037505-RA	
Dvex_051256-RA	XP_002122739.1	Dvex_051256-RA	88.6	4.1	sameW	scaffold69637-size2666	Dvex_051256-RA	
Dvex_007855-RA	XP_002123047.1	Dvex_007855-RA	77.6	1.11	invert	scaffold4244-size12665	Dvex_007855-RA	
Dvex_032347-RA	XP_002125803.1	Dvex_032347-RA	95.2	1	sameW	scaffold32319-size5462	Dvex_032347-RA	
Dvex_046616-RA	XP_002126036.1	Dvex_046616-RA	42.2	5.15	invert	scaffold58693-size3246	Dvex_046616-RA	
Dvex_022259-RA	XP_002126185.1	Dvex_022259-RA	92.8	1.42	invert	scaffold18411-size7497	Dvex_022259-RA	
Dvex_054175-RA	XP_002126765.1	Dvex_054175-RA	49.4	5.09	sameW	scaffold77562-size2314	Dvex_054175-RA	
Dvex_001129-RA	XP_002127117.1	Dvex_001129-RA	78.9	1.69	invert	scaffold329-size22474	Dvex_001129-RA	
Dvex_045777-RA	XP_002128363.1	Dvex_045777-RA	96.2	1.73	sameW	scaffold56828-size3360	Dvex_045777-RA	
Dvex_012313-RA	XP_002129163.1	Dvex_012313-RA	95.9	1.29	invert	scaffold7811-size10456	Dvex_012313-RA	
Dvex_028525-RA	XP_002129330.1	Dvex_028525-RA	88.4	1.41	sameW	scaffold26605-size6218	Dvex_028525-RA	
Dvex_005247-RA	XP_002129440.1	Dvex_005247-RA	99.8	1	invert	scaffold2541-size14540	Dvex_005247-RA	
Dvex_009309-RA	XP_002129488.1	Dvex_009309-RA	95.3	2.31	invert	scaffold5373-size11822	Dvex_009309-RA	
Dvex_056437-RA	XP_002130282.1	Dvex_056437-RA	85	5	invert	scaffold84291-size2048	Dvex_056437-RA	
Dvex_034023-RA	XP_002130303.1	Dvex_034023-RA	97.9	1.22	invert	scaffold34978-size5152	Dvex_034023-RA	
Dvex_036526-RA	XP_002130315.1	Dvex_036526-RA	98.2	2.75	sameW	scaffold39043-size4735	Dvex_036526-RA	
Dvex_014976-RA	XP_002131348.1	Dvex_014976-RA	48.3	5	sameW	scaffold10381-size9433	Dvex_014976-RA	
Dvex_013696-RA	XP_002131800.1	Dvex_013696-RA	83	1.23	invert	scaffold9124-size9883	Dvex_013696-RA	
Dvex_024840-RA	XP_002132074.1	Dvex_024840-RA	50	4.97	sameW	scaffold21620-size6951	Dvex_024840-RA	
Dvex_031979-RA	XP_004226046.1	Dvex_031979-RA	41.2	5.12	invert	scaffold31748-size5532	Dvex_031979-RA	
Dvex_007888-RA	XP_009858084.1	Dvex_007888-RA	96.6	1.03	sameW	scaffold4274-size12629	Dvex_007888-RA	
Dvex_036779-RA	XP_009858362.1	Dvex_036779-RA	77.6	3.85	sameW	scaffold39501-size4693	Dvex_036779-RA	
Dvex_014215-RA	XP_009858393.1	Dvex_014215-RA	88.2	1.49	invert	scaffold9635-size9689	Dvex_014215-RA	
Dvex_004206-RA	XP_009858812.1	Dvex_004206-RA	45.1	1.86	invert	scaffold1871-size15642	Dvex_004206-RA	
Dvex_000829-RA	XP_009860542.1	Dvex_000829-RA	89	1.02	sameW	scaffold209-size24088	Dvex_000829-RA	
Dvex_049037-RA	XP_009862116.1	Dvex_049037-RA	56.7	1.59	invert	scaffold64237-size2934	Dvex_049037-RA	
Dvex_022782-RA	XP_009862245.1	Dvex_022782-RA	74.9	3.29	sameW	scaffold19047-size7379	Dvex_022782-RA	
Dvex_056437-RA	XP_018669988.1	Dvex_056437-RA	86.2	4.87	sameW	scaffold84291-size2048	Dvex_056437-RA	
Dvex_007706-RA	XP_026691217.1	Dvex_007706-RA	85.7	2.42	sameW	scaffold4130-size12765	Dvex_007706-RA	
Dvex_001057-RA	XP_026691338.1	Dvex_001057-RA	96	1.51	sameW	scaffold306-size22841	Dvex_001057-RA	
Dvex_008320-RA	XP_026691551.1	Dvex_008320-RA	89.3	1.37	invert	scaffold4591-size12369	Dvex_008320-RA	

Ciona intestinalis	Inicio	Fin	Didemnum vexillum	Inicio	Fin
NC_020173.2	1295853	1296066	scaffold7811-size10456	7974	8187
NC_020173.2	1296482	1296700	scaffold7811-size10456	6063	6281
NC_020173.2	1296973	1297240	scaffold7811-size10456	5612	5879
NC_020173.2	1297668	1297882	scaffold7811-size10456	4930	5144
NC_020173.2	1298530	1298741	scaffold7811-size10456	4439	4650
NC_020173.2	1299243	1299426	scaffold7811-size10456	4128	4311
NC_020173.2	1299836	1300072	scaffold7811-size10456	3624	3860
NC_020173.2	1300289	1300479	scaffold7811-size10456	2896	3086
NC_020173.2	1300857	1301034	scaffold7811-size10456	2024	2201
NC_020173.2					

Cromosoma *C. intestinalis*	Inicio	Fin	Scaffold *D. vexillum*	Inicio	Fin
NC_020173.2	4974601	4974796	scaffold4274-size12629	4161	4356
NC_020173.2	4975054	4975340	scaffold4274-size12629	4640	4926
NC_020173.2	4975439	4976008	scaffold4274-size12629	5101	5670
NC_020173.2	4976202	4976482	scaffold4274-size12629	5880	6160
NC_020173.2	4976582	4976786	scaffold4274-size12629	6445	6649
NC_020173.2	4976741	4977059	scaffold4274-size12629	7253	7571

En la tabla: la Proteina del Fagosoma *C. intestinalis* 100178709, de la base de datos KEGG [KO:K07374] que, es Generada en el Gen LOC100178709(NCBI), LOCUS XP_009858084, con la version de proteina **XP_009858084.1;** Es ortóloga a la proteina *D. vexillum* Dvex_007888-RA, originada del scaffold "scaffold4274-size12629" y se identificaron Bloques Sinténicos Microsintenia con las regiones genomicas de *C. intestinalis* en el cromosoma 8(NC_020173.2).

Figura 2-15.: Identificación de microsintenia conservada entre *D.vexillum* y *C. intestinalis*

Identificación de Microsintenia entre *C. intestinalis, D. vexillum, C. savignyi, O.dioica* y el cefalocordado *B. floridae*

De manera similar al ejemplo ilustrado anteriormente, la identificación de microsintenia fue identificada entre *D. vexillum* y *C. intestinalis*, *C. savigyni*, *O. dioica*, *B. schlosseri* y el cefalocordado *B.floridae*; un primer paso la postulación de genes ortólogos y un orden conservada en las secuencias para posteriormente identificar la sintenia en las subregiones genómicas: se identificaron 4 regiones microsinténicas cuyos genes constitutivos se destacan de la tabla **2-17**,

Las gráficas que representan las regiones con microsintenia, muestra una baja conservación entre las especies *D. vexillum* y *C. intestinalis*, *C. savigyni*, *O. dioica*, *B. schlosseri* y el cefalocordado *B. floridae*. Finalmente, con el método propuesto de detección de proteínas ortólogas, cuyos resultados se presentan en la tabla **2-16** y la validación de las coordenadas de origen que conservan bloques de sintenia, se identificaron las relaciones sinténicas de las secuencias que se presentan en la tabla **2-17**. Para las relaciones de ortologías se detectaron 47, 42, 24 y 26 ortólogos entre *C. intestinalis* usado como referencia y *D.vexillum*, *C. savigni*, *B. floridae* y *O. dioca* respectivamente. Para las intersecciones de colinealidad un total de 37 relaciones entre *D.vexillum* y *C. intestinalis*, 13 con *C. savigni*, 26 con *B. floridae* y 32 con *O. dioca*. Con *B. schlosseri* dada su organización de mayor información genómica en un cromosoma undet, la cuantificación excedió las capacidades de cálculo en este trabajo.

C.intestinalis	*Dvexillum*	identidad	R.tamaño	S
NP_001027643.1	Dvex_036526-RA	50.4	2.8	-
NP_001027676.1	Dvex_034922-RA	57.2	5.23	-
NP_001027684.1	Dvex_021280-RA	62.1	5.14	+
NP_001027802.1	Dvex_046616-RA	43.5	3.58	+
NP_001028999.1	Dvex_007391-RA	54.8	1.03	-
NP_001029015.1	Dvex_012956-RA	81.5	4.03	-
NP_001121589.1	Dvex_016308-RB	68.4	4.01	-
NP_001121595.1	Dvex_042897-RA	70.9	4.24	+
XP_002121757.1	Dvex_037505-RA	77.8	1.95	+
XP_002122739.1	Dvex_051256-RA	88.6	4.1	+
XP_002123047.1	Dvex_007855-RA	77.6	1.11	-
XP_002125803.1	Dvex_032347-RA	95.2	1	+
XP_002126036.1	Dvex_046616-RA	42.2	5.15	-
XP_002126185.1	Dvex_022259-RA	92.8	1.42	-
XP_002126719.1	Dvex_054175-RA	50.5	4.91	+
XP_002126765.1	Dvex_054175-RA	49.4	5.09	+
XP_002127117.1	Dvex_001129-RA	78.9	1.69	-
XP_002128363.1	Dvex_045777-RA	96.2	1.73	+
XP_002129163.1	Dvex_012313-RA	95.9	1.29	-
XP_002129330.1	Dvex_028525-RA	88.4	1.41	+
XP_002129440.1	Dvex_005247-RA	99.8	1	-
XP_002129488.1	Dvex_009309-RA	95.3	2.31	-
XP_002130282.1	Dvex_056437-RA	85	5	-
XP_002130282.1	Dvex_056437-RA	85	5	-
XP_002130303.1	Dvex_034023-RA	97.9	1.22	-
XP_002130315.1	Dvex_036526-RA	98.2	2.75	+
XP_002130796.1	Dvex_001057-RA	43	3.88	-
XP_002131348.1	Dvex_014976-RA	48.3	5	+
XP_002131800.1	Dvex_013696-RA	83	1.23	-
XP_002132074.1	Dvex_024840-RA	50	4.97	+
XP_002132118.1	Dvex_049037-RA	72.9	1.59	+
XP_004226046.1	Dvex_031979-RA	41.2	5.12	-
XP_009858084.1	Dvex_007888-RA	96.6	1.03	+
XP_009858362.1	Dvex_036779-RA	77.6	3.85	+
XP_009858393.1	Dvex_014215-RA	88.2	1.49	-
XP_009858812.1	Dvex_004206-RA	45.1	1.86	-
XP_009860542.1	Dvex_000829-RA	89	1.02	+
XP_009862116.1	Dvex_049037-RA	56.7	1.59	-
XP_009862245.1	Dvex_022782-RA	74.9	3.29	+
XP_018667054.1	Dvex_007706-RA	43.8	2.44	+
XP_018669988.1	Dvex_056437-RA	86.2	4.87	+
XP_018669988.1	Dvex_056437-RA	86.2	4.87	+
XP_026690556.1	Dvex_016308-RB	81.9	4.04	-
XP_026691217.1	Dvex_007706-RA	85.7	2.42	+
XP_026691338.1	Dvex_001057-RA	96	1.51	+
XP_026691551.1	Dvex_008320-RA	89.3	1.37	-
XP_026696426.1	Dvex_051256-RA	40.9	3.5	+

C.intestinalis	*C. Savignyi*	identidad	R. Tamaño	S.
NP_001027676.1	ENSCSAVP00000005244.1	97.9	1	-
NP_001027684.1	ENSCSAVP00000019228.1	66.5	1.02	-
NP_001027787.1	ENSCSAVP00000020033.1	78	1.01	+
NP_001028999.1	ENSCSAVP00000008363.1	72.8	1.01	+
NP_001029008.1	ENSCSAVP00000009570.1	91.4	1	-
NP_001121588.1	ENSCSAVP00000014274.1	87.1	1.27	-
NP_001121589.1	ENSCSAVP00000019991.1	85.6	1.39	+
NP_001121595.1	ENSCSAVP00000000091.1	92.5	1.16	+
XP_002129488.1	ENSCSAVP00000004330.1	94.6	1	-
XP_002132074.1	ENSCSAVP00000007323.1	50.6	1.34	-
XP_002130230.1	ENSCSAVP00000017952.1	80	1.14	+
XP_026689431.1	ENSCSAVP00000016167.1	48.9	3.21	-
XP_002127117.1	ENSCSAVP00000015891.1	95.4	1	-
XP_002129330.1	ENSCSAVP00000009994.1	96.6	1	+
XP_002130303.1	ENSCSAVP00000014899.1	98.4	1	-
XP_002126765.1	ENSCSAVP00000016384.1	52.2	3.77	+
XP_002128363.1	ENSCSAVP00000014513.1	98.8	1	-
XP_002129025.1	ENSCSAVP00000008776.1	65.8	1	-
XP_002121757.1	ENSCSAVP00000004501.1	90.5	1	-
XP_002126036.1	ENSCSAVP00000019500.1	40.8	4.72	-
XP_009858084.1	ENSCSAVP00000016573.1	98.4	1	-
NP_001265897.1	ENSCSAVP00000002492.1	82.2	1.52	-
XP_002126159.1	ENSCSAVP00000005870.1	48.9	2.54	+
NP_001161178.1	ENSCSAVP00000008776.1	61.2	1.07	-
XP_002128393.1	ENSCSAVP00000005870.1	47.5	2.54	+
XP_002130282.1	ENSCSAVP00000017814.1	84.4	1	-
XP_026692805.1	ENSCSAVP00000003116.1	50.5	1.01	-
XP_009862245.1	ENSCSAVP00000005589.1	86.5	1.03	-
XP_002131800.1	ENSCSAVP00000001554.1	88.3	1.01	+
XP_002131348.1	ENSCSAVP00000007323.1	93.7	1	-
XP_002129962.1	ENSCSAVP00000008792.1	70.5	1	-
XP_002119312.1	ENSCSAVP00000004259.1	80.1	1.06	+
XP_002129440.1	ENSCSAVP00000019572.1	100	1	+
XP_026691356.1	ENSCSAVP00000010067.1	70.2	2.35	+
XP_002127919.1	ENSCSAVP00000010231.1	47.6	2.06	+
XP_002130315.1	ENSCSAVP00000015938.1	99.3	1	-
XP_002122148.1	ENSCSAVP00000007707.1	69.8	2.48	+
XP_002129163.1	ENSCSAVP00000012829.1	98	1.02	-
XP_002122739.1	ENSCSAVP00000003523.1	88	1	+
XP_002126185.1	ENSCSAVP00000003187.1	97	1.07	+
XP_018669988.1	ENSCSAVP00000017955.1	83.8	1.05	+
XP_004226046.1	ENSCSAVP00000002056.1	77.6	2.63	+

B. Floridae	*C. Intestinalis*	identidad	R. Tamaño	S.
XP_002592130.1	NP_001027676.1	92	1	-
XP_002594487.1	NP_001041458.1	42.2	5.15	-
XP_002612231.1	NP_001121595.1	67.1	1.01	+
XP_002586964.1	XP_002129488.1	68.6	1.03	-
XP_002598155.1	XP_002132074.1	81.6	1.03	-
XP_002610712.1	XP_026689431.1	75.8	2.28	-
XP_002610640.1	XP_002127117.1	76.5	1.02	-
XP_002600269.1	XP_002129330.1	77.8	1.05	+
XP_002596422.1	XP_002126765.1	66.7	1	-
XP_002602796.1	XP_002128363.1	91.2	1.15	+
XP_002600490.1	XP_002129025.1	56.7	3.54	-
XP_002609521.1	XP_002121757.1	65.7	1	+
XP_002595073.1	NP_001265897.1	53.5	1.95	+
XP_002591668.1	NP_001161178.1	46.8	1.62	+
XP_002593756.1	XP_026692805.1	48.2	1.91	+
XP_002590418.1	XP_009862245.1	57.6	1.76	-
XP_002606860.1	XP_002131800.1	72	1	+
XP_002610356.1	XP_002131348.1	67.9	1	-
XP_002608917.1	XP_002130315.1	98.9	1	+
XP_002593592.1	XP_002123047.1	63.5	1.11	-
XP_002609084.1	XP_002129163.1	89.1	1	+
XP_002587657.1	XP_026692891.1	50.2	4.5	+
XP_002605275.1	XP_002126185.1	80.6	1	+
XP_002611753.1	XP_018669988.1	69.2	1.08	+

C.intestinalis	*O. Dioica*	identidad	R. Tamaño	S.
NP_001027676.1	CBY36918.1	79.3	1.01	+
NP_001028999.1	CBY41582.1	41	2.11	-
NP_001041458.1	CBY40826.1	43.2	5	+
NP_001161178.1	CBY41871.1	45.7	5.02	+
XP_002119312.1	CBY38381.1	44.5	4.86	-
XP_002121757.1	CBY41557.1	52.2	1.32	-
XP_002122739.1	CBY43752.1	85	1.01	-
XP_002123047.1	CBY39336.1	58.6	1.15	+
XP_002126036.1	CBY39130.1	44.7	1.23	+
XP_002126159.1	CBY38876.1	45.5	3.54	-
XP_002126185.1	CBY32629.1	41.3	4.13	+
XP_002126765.1	CBY32870.1	55.9	1.01	-
XP_002127117.1	CBY33512.1	71.1	1.04	-
XP_002128363.1	CBY40969.1	87.6	1.76	+
XP_002128393.1	CBY38876.1	45.8	3.54	-
XP_002129163.1	CBY39635.1	82.5	3.57	-
XP_002129330.1	CBY40548.1	65.6	1.04	+
XP_002129440.1	CBY43752.1	98.9	1	-
XP_002129488.1	CBY38348.1	47.3	3.34	-
XP_002129962.1	CBY42127.1	43.3	4.72	+
XP_002130282.1	CBY37572.1	60.9	3.18	+
XP_002130303.1	CBY43752.1	97.1	1	-
XP_002130315.1	CBY41790.1	98.6	1	-
XP_002131348.1	CBY38608.1	47.5	1.31	-
XP_002131800.1	CBY30985.1	70.1	1.07	+
XP_002132074.1	CBY38608.1	79.2	1.01	-

Figura 2-16.: Resumen de candidatos a genes ortólogos del fagosoma, entre *C.intestinalis*, *D.vexillum*, *C. intestinalis* y el cefalocordado *B. floridae*

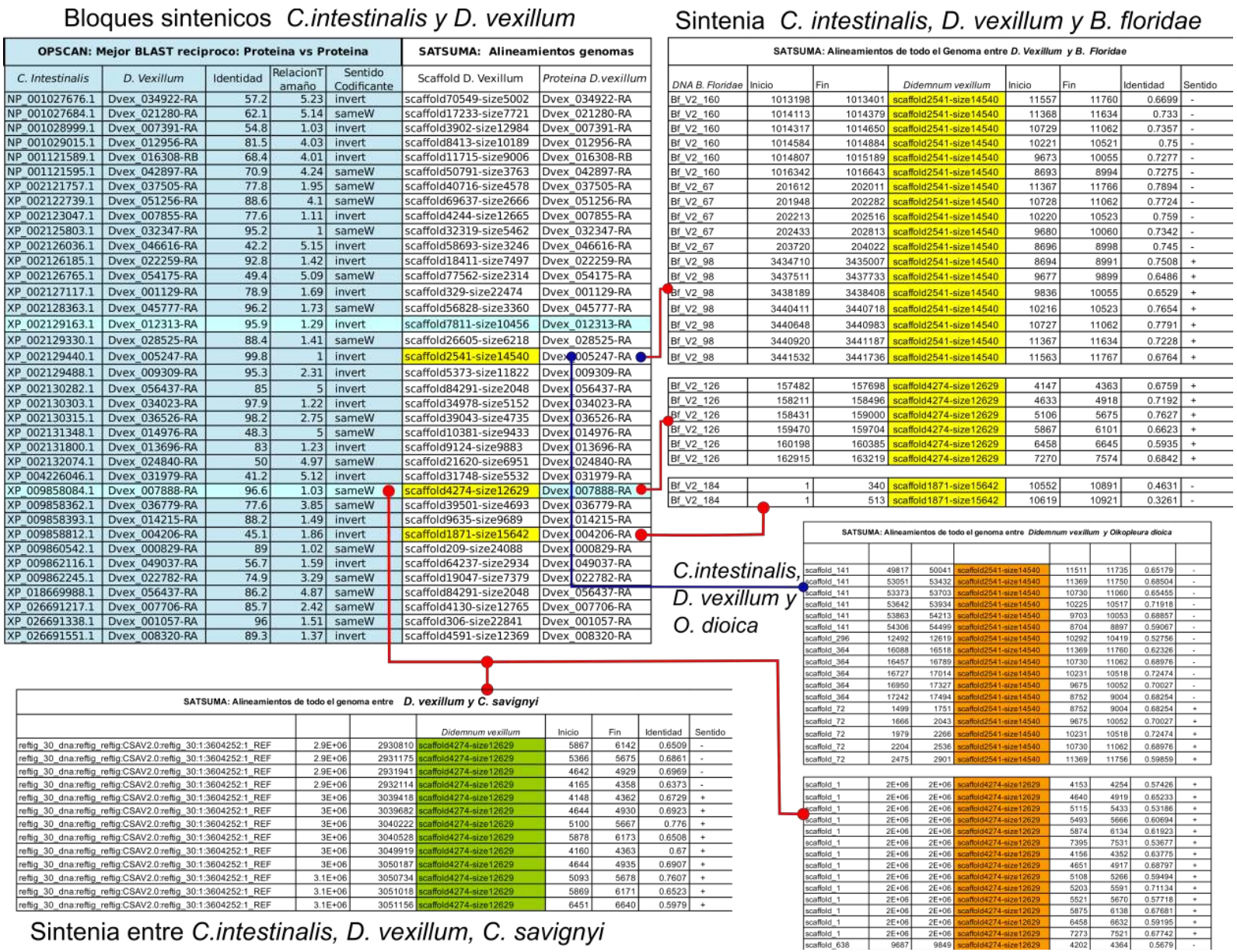


Figura 2-17.: Identificación de sintenia entre las especies *C. intestinalis*, *C. savigyni*, *O. dioica*, *B. schlosseri* y el cefalocordado *Branchiostoma floridae*.

3. Visualización de la anotación funcional y estructural del genoma de *D .vexillum* y regiones asociadas al sistema inmune siguiendo un enfoque sinténico

En décadas pasadas los proyectos de investigacion genomica generó enormes cantidades de datos como se resume en ver gráfica **3-1** [1] que son almacenados en sitios centralizados en donde se relacionan y son presentados de manera clara; convirtíendo estas bases de datos en recursos esenciales y de uso habitual por los bioólogos en todo el mundo[Stein, 2003]; ellas son una via rápida para disponer de una gran cantidad de información desde la que los investigadores puedan interpretar sus observaciones, diseñar nuevos experimentos y obtener información sobre la estructura génica, la organización del genoma y su evolución [Sensen, 2005].

El diseño de la base de datos y el desarrollo del software es un factor limitante para construir nuevas bases de datos y las existentes no puede ser simplemente copiadas para crear bases de datos de otros organismos, sin embargo, los investigadores podrían desarrollar nuevas bases de datos para sus organismos aprovechando el trabajo hecho sobre las que ya existen, de otro modo, implica una innecesaria duplicacion de recursos. Desafortunadamente, cuando en las investigaciones se desarrollan metodos, independientemente de cual sea el tipo de datos, no es seguro que la forma en que se almacenan e intercambian los datos genómicos tenga exactamente el mismo significado que el que le otorgan otras bases de datos biológicas independientes. Estas diferencias en la definicion de datos podrıa dificultar la consulta entre las bases de datos[NIGMS, 2002]: Para que exista un crecimiento en las bases de datos para nuevos organismos se requiere una reducción en los costos de las Tecnologías de información al proveer un desarrollo de software y el diseño de la base de datos con un estándar de almacenamiento e intercambio de los datos biológicos con otras bases de datos ya existentes, estas son las razones por las que existe el proyecto GMOD y el esquema relacional Chado.

Chado es un esquema relacional para Bases de Datos mediado por Ontologías que es capaz de

[1] Imagen tomada de `https://www.ncbi.nlm.nih.gov/refseq/statistics` en Mayo de 2020

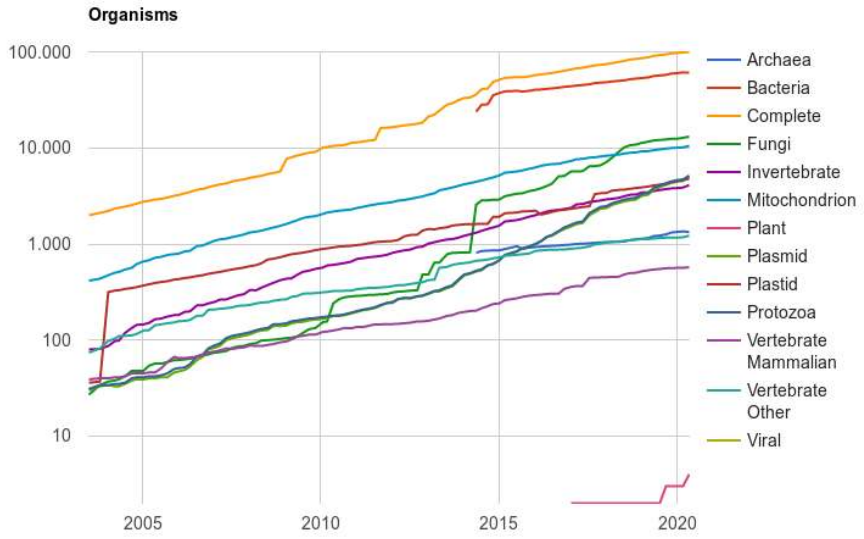


Figura 3-1.: Crecimiento de los organismos durante última década.

representar muchos de los tipos de datos encontrados en la Biología moderna [Mungall *et al.*, 2007]: una ontología es esencialmente un diccionario de términos con un vocabulario controlado que, en chado tiene dos componentes; los términos de ontología(anotaciones GO) que se usan para etiquetar las secuencias caracterizadas(p. eje. exon, UTR, gene), así como para etiquetar el origen de la misma secuencia (p. eje. para predicción computacional Maker, snap, augustus) y las ontologías GO con las que se definen el tipo de relaciones que existen entre las secuencias etiquetadas: Mientras una secuencia caracterizada se etiqueta con términos únicos con la que es descrita(p. eje. tipo, origen), las relaciones entre las secuencias etiquetadas pueden tener multiples términos que las asocien, p.eje. (parte de...es un). La figura **3-6** muestra un grafo[2] que representa las relaciones de ontología donde, los vértices representan los términos de ontología y las aristas las relaciones entre estos términos.

3.1. Proyecto GMOD: Una base de datos para Organismos modelo

El esquema relacional de bases de datos Chado es parte de GMOD –acrónimo de Generic Model Organism Database– que es un proyecto Open Source cuyo objetivo es desarrollar una suite de software para crear y administrar Bases de Datos de Organismos Modelo. Los componentes de este proyecto incluyen herramientas de visualización y edición del genoma, Herramientas de Curación, herramientas para Ontologías Biólogicas y un set de procedimientos operativos (rutinas computacionales), entre otras. El proyecto es fundado

[2]Tomado de `http://gmod.org/wiki/Introduction_to_Chado`

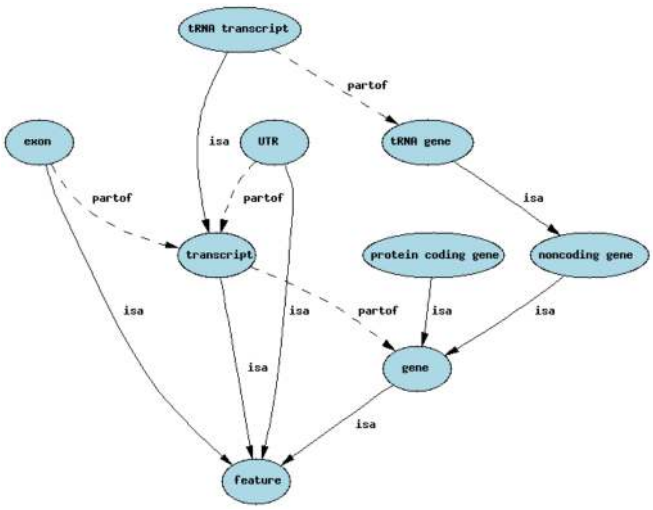


Figura 3-2.: Grafo representando algunas relaciones de ontologia en chado

por el NIH (http://www.nih.gov/), la Dependencia "Servicio de investigación Agrícola" del Departamento de Agricultura de Estado Unidos "USDA" (http://www.ars.usda.gov/) y la Fundación Nacional de Ciencia (http://www.nsf.gov/). Tambien se cuenta como participes a miembros de proyectos de bases de datos incluyendo WormBase (http://wormbase.org/), FlyBase (http://flybase.org/), Mouse Genome Informatics (http://www.informatics.jax.org/), Gramene (http://gramene.org/), the Rat Genome Database (http://rgd.mcw.edu/), TAIR (http://arabidopsis.org/), EcoCyc (http://ecocyc.org/), DictyBase (http://dictybase.org/), wFleaBase (http://wfleabase.org/), NESCent (http://www.nescent.org/), y Saccharomyces Genome Database (http://yeastgenome.org/).

Entre las herramientas incluidas en GMOD estan[3] GBrowse [Stein *et al.*, 2002], JBrowse [Skinner *et al.*, 2009], CMap [Youens-Clark *et al.*, 2009], Pathway Tools [Karp *et al.*, 2010], Sybil [Crabtree *et al.*, 2007], Apollo [Lee *et al.*, 2009], BioMart [Kasprzyk, 2011], InterMine [Smith *et al.*, 2012], Maker [Campbell *et al.*, 2014], Tripal [Sanderson *et al.*, 2013] , Galaxy [Afgan *et al.*, 2018].

3.1.1. Instalación de la base de datos PostgreSQL y el esquema relacional Chado

Como enfoque general el servidor web recibe las peticiones desde la aplicación que esta usando el usuario y las traduce a consultas para la base de datos. La base de datos recibe la petición y entrega los datos que son gestionados con otras aplicaciones que median y que usando formatos los envía de nuevo a la aplicación para que ésta la presente al usuario.

[3] `http://gmod.org/wiki/GMOD_Components`

La base de datos instalada sigue un modelo GMOD[4], por lo que tiene una arquitectura general en común:

- Una Base se Datos.
- Un flujo de programas que median entre la Aplicación y la Base de Datos.
- Un conjunto de Aplicaciones WEB más cercanas al usuario.

El montaje del Gestor de Base de Datos PostgresSQL[5] además de centralizar la administración de los datos logra:

1. Dar un manejo uniforme a todos los datos obtenidos en la fase de modelamiento y anotación
2. Un eficiente uso del lenguajes de computadora para el almacenamiento, acceso, actualización y manejo de los datos.
3. Lograr una descripción de todos los objetos y estructuras de la base de datos: Metadatos
4. Lograr una relación usuario/aplicación-visualización especificas a cada usuario y aplicación.
5. Ajustar las limitaciones de integridad con los datos
6. Implementar mecanismos de seguridad para proteger los datos.
7. Lograr un conjunto combinado de transacciones y operaciones sobre datos dentro de unidades atómicas.
8. Lograr una sincronización de transacciones para los usuarios actuales.
9. Lograr la recuperación de datos luego de la caída del sistema
10. Manejar consultas agregadas, generacion de reportes, interfaces para otras bases de datos e interfaces para otras aplicaciones.

La instalación de la base de datos PostgreSQL y el esquema relacional Chado se realizó de acuerdo a la descripcion de Scott Cain[6][7]. Se usó la versión 8.1.8 del sistema gestor de base de batos PostgreSQL. Los módulos PERL fueron instalados desde la línea de comandos con CPAN[8].
La version de Chado instalada corresponde a la version 1.70, en ella se incluyeron las ontologías de *Relationship Ontology*, *Sequence Ontology*, *Gene Ontology,Chado Feature Properties, Cell Ontology* y *Plant Ontology*.

[4] `http://gmod.org/wiki/Overview`
[5] `https://www.postgresql.org/`
[6] `http://gmod.org/wiki/User:Scott`, consultado en 14 de julio de 2020
[7] `http://gmod.org/wiki/Chado_-_Getting_Started#Installation`, consultado en 14 de julio de 2020
[8] `http://www.cpan.org/`, consultado en 14 de julio de 2020

3.1.2. Poblado de registros en la base de datos

El poblado de datos en la base se hizo según `http://gmod.org/wiki/Load_GFF_Into_Chado` se usa el *script* en PERL *gmod_ bulk_ load_ gff3.pl* para poblar la base de datos con los registros seleccionados de la etapa anterior y a los cuales BLAST2GO ha hecho anotación funcional con ontologías.

3.2. Montaje del FrontEnd para visualizar Datos Genómicos de D. vexillum

Además de crear la base de datos y poblarla con los resultados de datos genómicos de *D. vexillum*, es preferible incluir un diseño de página web y entornos gráficos de facil uso. Toda la funcionalidad que puede describir un sitio web para un organismo requiere mucho tiempo y es muy costosa[O'Connor *et al.*, 2008], por lo que podrían llegar a ser prohibitivas este tipo de herramientas. Como resultado de este panorama, en la actualidad para el proyecto GMOD existen dos herramientas que cubren esta necesidad; GMODWeb y TRIPAL[9].
Nosotros usamos GMODWeb construido con lenguaje PERL utilizado en la herramienta de software, Turnkey[10]: La idea[11] del proyecto Turnkey es que se puede tomar el esquema de cualquier base de datos y con el grupo de modulos SQLFairy[12] de PERL convertirlo en objetos[13] que representen tablas y columnas; la salida estos objetos y las relaciones inferidas desde el modulo SQL::Translator, son ajustadas en plantillas Basadas en otro módulo PERL Template Toolkit[Darren *et al.*, 2004][14] representando los objetos del esquema. Todo el proceso sigue un modelo llamado MVC: una Base de Datos Modelada(Model), Visualizada(View) y Controlada(Controller) con elementos de diseño Turnkey[15]. GMODWEB automaticamente creará un website interpretable con el uso del módulo perl para Apache2.
Existen varias ventajas[16] al usar Turnkey:

1. Turnkey puede tomar un buen diseño de esquema relacional SQL y crear un Web Site totalmente funcional en pocos minutos, este sitio puede ser usado para buscar relaciones de tablas, buscar registros particulares, ver los vínculos entre registros y actualizar los valores de los campos.

[9] `http://gmod.org/wiki/Tripal`
[10] http://radius.genomics.ctrl.ucla.edu/turnkey/pmwiki.php?n=Main.HomePage
[11] http://radius.genomics.ctrl.ucla.edu/turnkey/pmwiki.php?n=Main.About
[12] `http://search.cpan.org/~jrobinson/SQL-Translator-0.11007/lib/SQL/Translator.pm`SQLFairy es tambien conocido como SQL::Translator
[13] Transforma los archivos SQL DDL en otra variedad de formatos, incluyendo otros dialectos SQL, documentación, imagenes y codigo
[14] http://search.cpan.org/ abw/Template-Toolkit-2.22/lib/Template/Tutorial/Web.pod
[15] Turnkey es un producto de SQLFairy, identificable por el tipo de datos visualizados en la salida, como GraphViz o Mysql
[16] `http://radius.genomics.ctrl.ucla.edu/turnkey/pmwiki.php?n=Main.UseCases`

2. Turnkey puede autogenerar una vista modelada y controlada desde el esquema relacional de base de datos. Una buena característica es que los componentes de manera individual pueden ser usados para construir aplicaciones más complejas. La capa del Objeto–Relacional puede ser usada en otras aplicaciones y así acceder a nivel de la base de datos. Adicionalmente, las visualizaciones se basan en una plantilla Perl del módulo Template::Toolkit[Darren *et al.*, 2004]: Esto permite planear nuevos plantillas para funcionalidades adicionales la aplicación web, soporta la personalización de fondos y vistas por parte del usuario. dinámicamente.

3.3. Resultados

La base de datos se presenta al usuario como un entorno web,

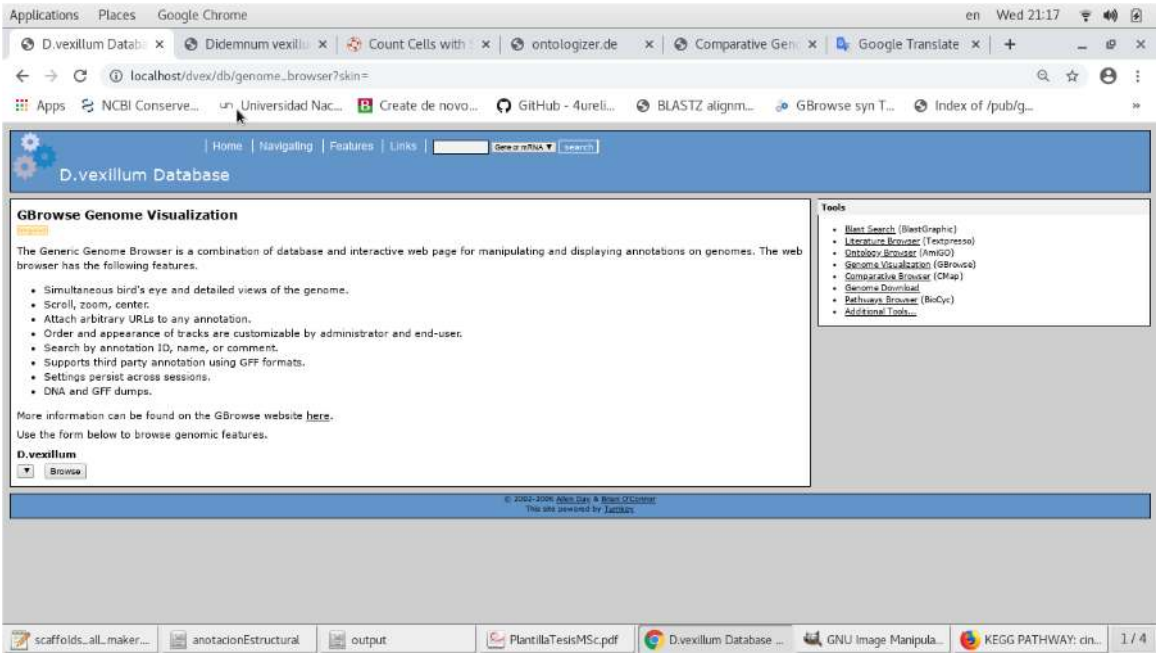


Figura 3-3.: FronEnd de inicio de la Base de datos para *vexilum*

Se pueden hacer búsquedas simples o indexadas:
La base de datos se presenta al usuario como un entorno web,

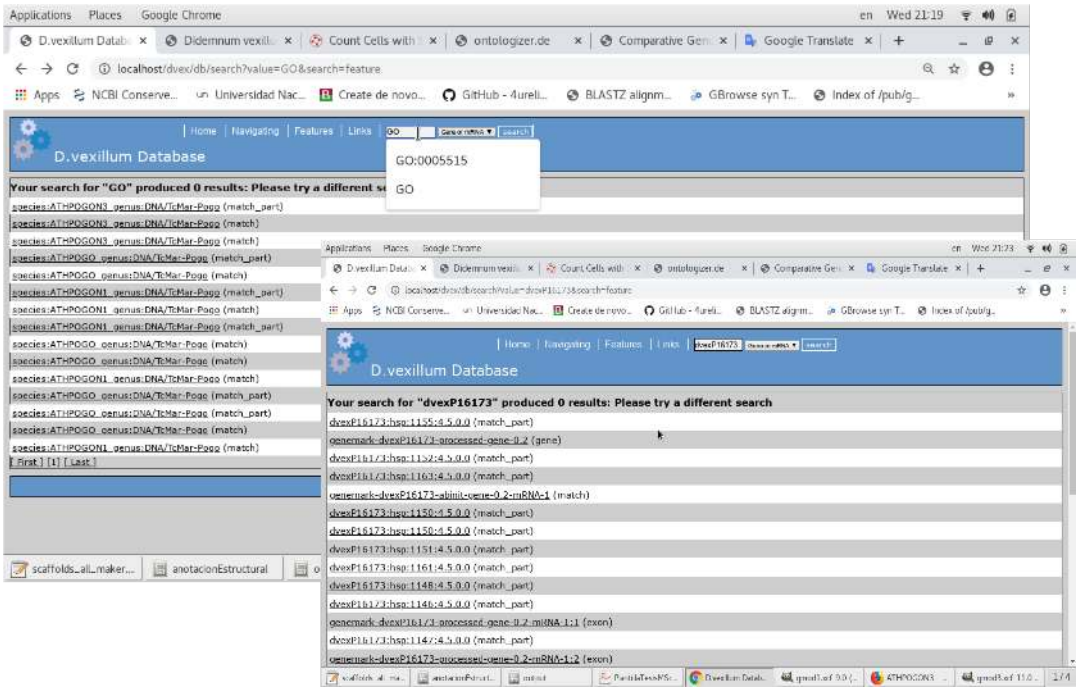


Figura 3-4.: FronEnd bienvenida de la base de datos para *D. vexilum*

Usa el navegador genómico Gbrowse:

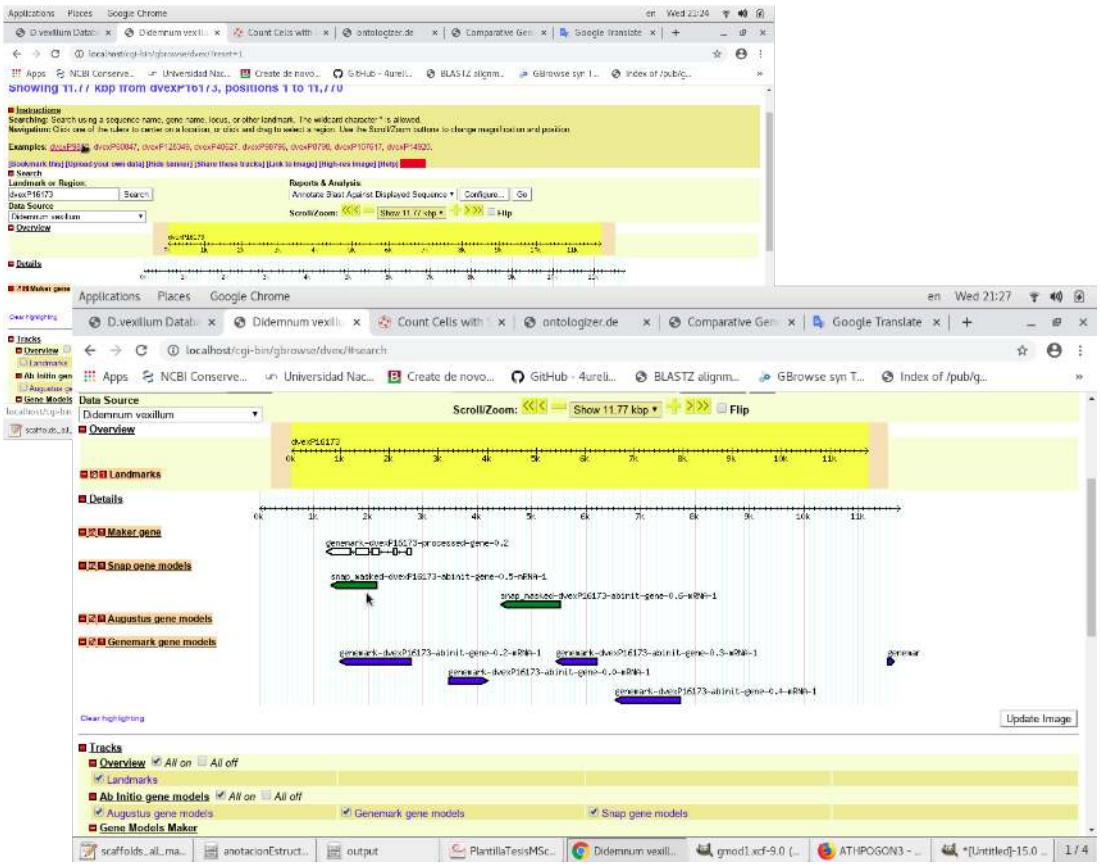


Figura 3-5.: Permite búsquedas simples sobre los resultados obtenidos de *D. vexilum* en este trabajo.

Permite mostrar el soporte experimental de los resultados obtenidos.

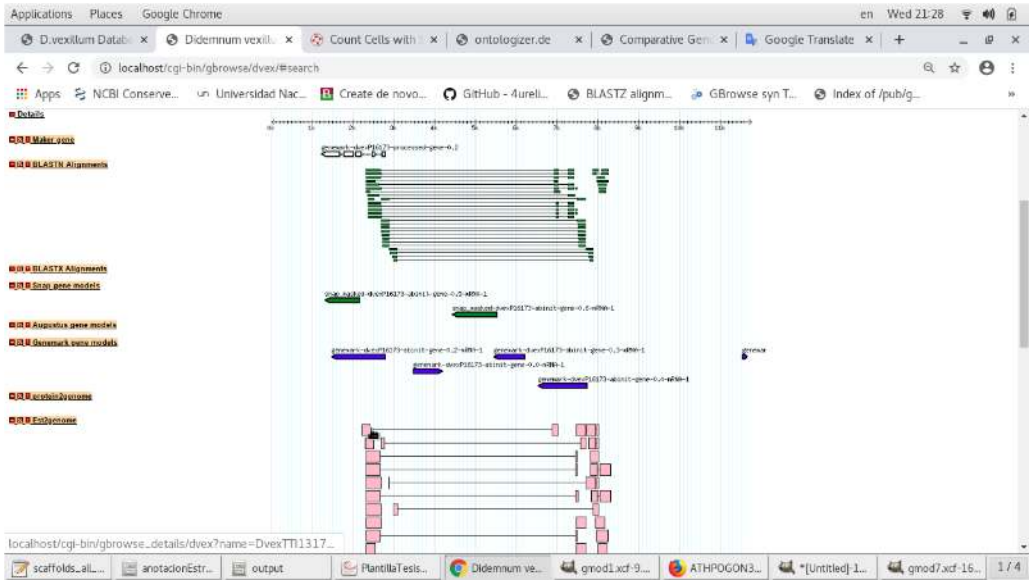


Figura 3-6.: El navegador genómico muestra la evidencia experimental de los resultados de *D. vexillum*

Como resultado final de este capítulo se tiene que se implementó la base de datos la información genómica procesada en esta tesis para la especie *D. vexillum*, en una plataforma web que permite realizar búsquedas sencillas e indexadas y donde los registros son relacionados usando Ontologias Biológicas para servicio de la comunidad científica. La base de datos es un trabajo que puede ser extendido a la administración y manejo del registro de otros organismos, procariotas-eucariotas de los que se logra obtener un texto plano GFF3 que describa sus características, estructurales y funcionales.

4. Conclusiones

- En el presente trabajo se combinaron los protocolos existentes y ajustados para re-ensamblar de forma híbrida el genoma de la especie *D. vexillum* pese a las dificultades experimentales del proceso de aislamiento y secuenciamiento del DNA genómico. Se obtuvo un ensamble genómico de un tamaño similar al reportado por Velandia-Huerto en 2016. Sin embargo, este nuevo ensamble presenta baja redundancia, y scaffolds con N50 cercano a ocho veces la longitud con respecto al primer reporte. Un ensamble híbrido con lecturas PacBio e Illumina como el obtenido en este trabajo, logró aumentar la continuidad en las lecturas, mientras que el error fue solucionado por la profundidad y bajo error soportada por las lecturas cortas de Illumina

- El uso de QUIVER como herramienta de pulido en este ensamble híbrido logra obtener secuencias, con una fiabilidad $\geq 99.9\,\%$. Sin embargo esta herramienta de pulido está limitada al uso de lecturas PacBio, por lo que se evidencia una menor longitud en las secuencias, las faltantes secuencias pertenecen a secuencias Illumina.

- El tamaño en scaffold en el nuevo ensamble genómico permitió la anotación estructural y funcional de *D. vexillum* que, junto con el ensamble de transcritos posibilitó identificar 90.938 transcritos correspondientes a 62194 genes con tamaño promedio de 1736Kb y con soporte experimental de su transcrito.

- Se postuló candidatos a proteínas codificantes de *D. vexillum* que amplia y complementa los trabajos más relevantes a nivel genómico que hasta el momento han sido encaminados a la identificación molecular de ncRNAs. En un alto porcentaje de las proteínas modeladas se anotó por trasferencia de homología a las proteínas de los tunicados *C. intestinalis* y *C. savignyi*. Se realizó una anotación funcional con UNIREF90 y PFAM, seguida de una clasificación de secuencias con ontologías asociadas al sistema inmune.

- A nivel de proteina la homología entre 81 proteinas del fagosoma de *C. intestinalis* y *D. vexillum* muestra que un 86 % de las proteínas se conservan con un cubrimiento promedio del 70 % y un e-value 1E-25.

- Se identificaron 42 proteinas candidatas ortólogas en *D. vexillum* que fueron asociadas al fagosoma de *C. intestinalis* . A nivel de nucleotido, *C. intestinalis*, *C. savigyni*, *O. dioica*, *B. schlosseri* y el cefalocordado *B. floridae* presenta una alta divergencia y

sólo aproximadamente entre el 20 y 30 % de secuencias del fagosoma de *C. intestinalis* pudieron compartir relación de ortología.

- Se identificaron 4 regiones genómicas que conservan microsintenia que fue asociada al sistema inmune entre *C. intestinalis* y *D. vexillum*. Sin embargo, no todas las regiones de microsintenia identificadas tienen origen en coordenadas genómicas de *C. intestinalis*, las otras regiones que presentan microsintenia fueron identificadas entre: *D. vexillum*, *C. savigyni*, *O. dioica* y el cefalocordado *B. floridae*.

- Los resultados preliminares de bloques sinténicos computados para las especies de estudio podrán en un futuro ser utilizados en aproximaciones actuales del refinamiento del proceso de scaffolding del genoma de la misma especie y permitirán mejorar en un futuro los procesos de anotación del los genomas de los tunicados y en particular de la especie estudiada.

- Toda la información de este trabajo quedó a disposición de la comunidad científica: los registros fueron organizados en una base de datos que usa el esquema relacional Chado, el navegador genómico Gbrowse y el buscador de homología Blast, en un entorno web que permite acceder a este recurso biológico.

- El trabajo realizado al construir la base de datos posibilita usar la plataforma por otro eucariote o procariote para evidenciar y colocar a disposicion todo trabajo con sentido biológico.

5. Recomendaciones y Productos

5.1. Recomendaciones

- Se podría mejorar la detección de regiones sinténicas incorporando nuevas reensamblajes genómicos que aprovechen la ventaja de producir lecturas largas para poder incrementar el L50 como Oxford Nanopore Technology y algoritmos que mejoran la detección sinténica como SynChro para la reconstrucción y visualización de los bloques sinténicos
- Para mejorar la anotación y validación de genes se sugiere trabajar con diferentes estadios del desarrollo de las colonias de la especie para obtener el mayor complemento de transcritos expresados diferencialmente ya que la actual información sólo captura la información expresada en el estado de vida de la colonia utilizada como estudio.

5.2. Productos

La información genómica y trasncriptómica generada en esta tesis es parte de la publicación sometida a evaluación en la revista BMC genomics

RESEARCH

The genome of the "Sea Vomit" *Didemnum vexillum*

Ernesto Parra-Rincón[1†], Cristian A. Velandia-Huerto[7,1†], Jörg Fallmann[7], Adriaan Gittenberger[2,3,4], Federico D. Brown[5,6], Peter F. Stadler[7,8,9,1,10] and Clara I. Bermúdez-Santana[1*]

Abstract

Background: Tunicates are the sister group of vertebrates and thus occupy a key position for investigations into vertebrate innovations as well as into the consequences of the vertebrate-specific genome duplications. Nevertheless, tunicate genomes have not been studied extensively in the past and comparative studies of tunicate genomes have remained scarce. The carpet sea squirt *Didemnum vexillum* is a colonial tunicate considered an invasive species with substantial ecological and economical risk.

Results: We report a newly re-assembled genome of *Didemnum vexillum*. We used a hybrid approach that combines two genome sequencing technologies and also its first transcriptome. Started from 28.5 Gb Illumina and 12.35 Gb of PacBio data a new hybrid scaffolded assembly was obtained comprised of a total size of 517.55 Mb that increases contig length about 8-fold compared to previous, Illumina-only assembly. While still highly fragmented (L50=25,284, N50=6539), the assembly is sufficient for comprehensive annotations of both protein-coding genes and non-coding RNAs.

Conclusions: The draft assembly of the "sea vomit" genome provides a valuable resource for comparative tunicate genomics and for the study of the specific properties of colonial ascidians.

Availability: Genome and annotation data as well as a link to a UCSC Genome Browser hub are available at `http://tunicatadvexillum.bioinf.uni-leipzig.de/`.

Keywords: Tunicata; *Didemnum vexillum*; microRNAs; genome annotation

Background

The carpet sea squirt *Didemnum vexillum* [1], a.k.a. "sea vomit", is a colonial tunicate presumably native to Japan that has appeared as an invasive species in Europe, the Americas, and New Zealand [2]. It negatively affects established benthic species and damages ship hulls as well as the infrastructure in marinas, ports, and shellfish farms.

Rapid colony growth or regression in response to the dynamics of the habitat [3], water temperature [4], colony fragmentation as a reproductive and dispersal strategy [5], fast asexual budding that allows attachment to a variety of living and/or non-living substrata, and relatively few predators [3] have facilitated *D. vexillum* to become a well-recognized world-wide invasor.

*Correspondence: cibermudezs@unal.edu.co
[1]Biology Department, Universidad Nacional de Colombia, Carrera 45 # 26-85, Edif. Uriel Gutiérrez, Bogotá D.C, Colombia
Full list of author information is available at the end of the article
†Equal contributor

The invasion potential of *D. vexillum* has an important economic impact on the aquaculture industry as it affects the conditions of bivalve and shellfish cultures (see e.g. [6] and the references therein), and increases the cost of maintenance to avoid the fouling process on mussel cages and facilities [7].

Despite the economic impact of tunicates and their pivotal phylogenetic position as sister group of the vertebrates, genomic studies and comparative analyses have remained relatively scarce. So far, the genomes of three solitary tunicates have been assembled and annotated in substantial depth: for the closely related sessile ascidians *Ciona savignyi* and *Ciona robusta* assemblies of their 14 chromosomes are available [8–11]; and for the pelagic larvacean *Oikopleura dioica* only 6 chromosomes have been reported [12–14]. In addition, draft assemblies recently have become available for the pelagic colonial thaliacian *Salpa thompsoni*, which was used to analyse the high mutation rates in the genomes of tunicates [15].

A. Electroferograma de Control de Calidad a las muestra biológicas Dvex2 y Dvex3 de *D. vexillum*

Overall Results for sample 7 : **D.vex2_CS.sample**

Number of peaks found:	3	Area 2:	66.8
Area 1:	165.5		

Peak table for sample 7 : **D.vex2_CS.sample**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	50	8.30	251.5	Lower Marker
2	2,050	3.31	2.4	
3	2,686	3.01	1.7	
4	4,748	1.23	0.4	
5	17,000	4.20	0.4	Upper Marker

Region table for sample 7 : **D.vex2_CS.sample**

From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
1,500	7,000	165.5	41	3,630	43.7	9.72	
7,000	35,000	66.8	16	15,029	32.3	3.71	

Assay Class: DNA 12000
Data Path: C:\...51228_pcb_km_c_DNA 12000_DE13805486_2015-12-28_10-17-08.xad

Created: 12/28/2015 10:17:08 AM
Modified: 12/28/2015 10:56:35 AM

Electropherogram Summary Continued ...

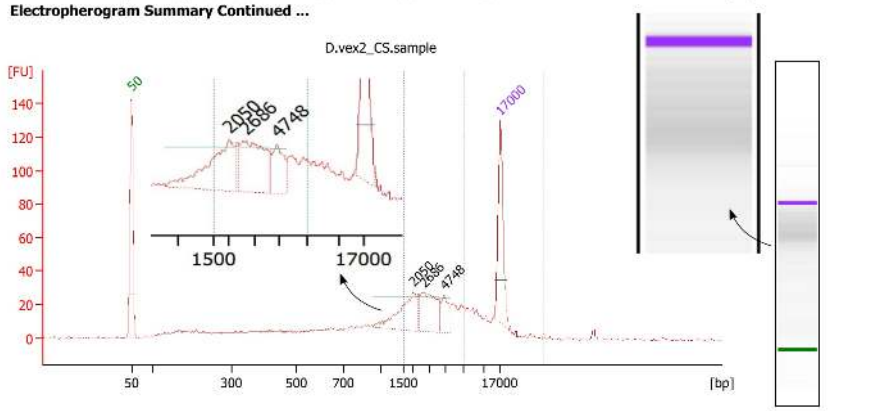


Figura A-1.: Electroferograma tomado a la muestra "Dvex2" antes de ser ligada a los adaptadores SmartBell de Pacbio; este análisis de calidad cuantificó la distribución y concentración de fragmentos de DNA del organismo *D.vexillum* antes de su secuenciamiento por Técnica Pacbio

Overall Results for sample 8 : D.vex3_CS.sample

Number of peaks found:	0	Area 2:	50.8
Area 1:	131.9		

Peak table for sample 8 : D.vex3_CS.sample

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	50	8.30	251.5	Lower Marker
2	17,000	4.20	0.4	Upper Marker

Region table for sample 8 : D.vex3_CS.sample

From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
1,500	7,000	131.9	42	3,739	42.6	5.57	
7,000	35,000	50.8	16	15,257	28.6	2.04	

Assay Class: DNA 12000
Data Path: C:\...51228_pcb_km_c_DNA 12000_DE13805486_2015-12-28_10-17-08.xad

Created: 12/28/2015 10:17:08 AM
Modified: 12/28/2015 10:56:35 AM

Electropherogram Summary Continued ...

Figura A-2.: Electroferograma tomado a la Muestra Dvex3 antes de ser ligada a los adaptadores SmartBell de Pacbio; este analisis de calidad cuantifico la distribucion y concentración de fragmentos de DNA del organismo D.vexillum antes de su secuenciamiento por Tecnica Pacbio

B. Creacion de Alias a los nombres PacBio

El secuenciamiento PacBio asigna a los resultados de cada celda una nomenclatura descrita en la figura **B-1**, para comodidad en la descripción se asignaron los alias que son relacionados en la tabla B

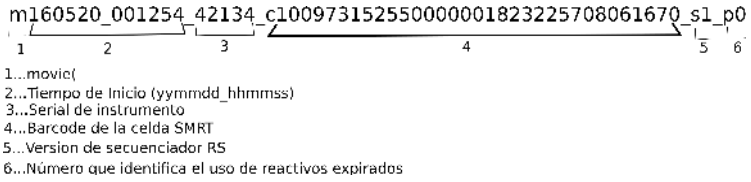


Figura B-1.: Seis primeros campos de la nomenclatura generados desde el secuenciamiento del gDNA del organismo D. vexillum

PacBio Name	Alias	Insert Preparation
m160520_001254_42134_c100973152550000001823225708061670_s1_p0	Cell_1	SS_10kb
m160617_012255_42134_c101009522550000001823230710211670_s1_p0	Cell_2	SS_10kb
m160617_074023_42134_c101009522550000001823230710211671_s1_p0	Cell_3	SS_10kb
m160617_135945_42134_c101009522550000001823230710211672_s1_p0	Cell_4	SS_10kb
m160720_123515_42134_c101034562550000001823250511171692_s1_p0	Cell_5	3_5kb
m160822_182901_42134_c101094012550000001823265402241725_s1_p0	Cell_6	3_5kb
m160822_224805_42134_c101094012550000001823265402241726_s1_p0	Cell_7	3_5kb
m160823_030725_42134_c101094012550000001823265402241727_s1_p0	Cell_8	3_5kb
m160823_072948_42134_c101090192550000001823245003091770_s1_p0	Cell_9	3_5kb
m160823_114551_42134_c101090192550000001823245003091771_s1_p0	Cell_10	3_5kb
m160929_075728_42134_c101080472550000001823240502241797_s1_p0	Cell_11	3_5kb
m160406_014229_42134_c100991432550000001823221607191666_s1_p0	Cell_12	20kb
m160416_021310_42134_c100991652550000001823221607191680_s1_p0	Cell_13	20kb
m160426_011156_42134_c101000102550000001823229908031615_s1_p0	Cell_14	20kb

Tabla B-1.: Relación de alias con la extensa nomenclatura de nombres en Pacbio

C. Caracterizacion de las Lecturas Cortas del secuenciamiento Illumina de ADNg de *D. vexillum*

C.1. Resumen cuantitativo

		GC	GC stdev	A	C	G	T	Millones Lecturas	N90	N50	Longitud Total
Run1	Forward	0.37	0.09	0.32	0.17	0.19	0.30	35,1	75	75	2,6Gbases
	Reverse	0.37	0.10	0.30	0.18	0.18	0.31	35,1	75	75	2,6Gbases
		GC	GC stdev	A	C	G	T	Millones Lecturas	N90	N50	Longitud Total
Run3	Forward	0.39	0.09	0.30	0.19	0.20	0.29	38.3	151	151	5.6Gbases
	Reverse	0.39	0.08	0.30	0.19	0.20	0.30	37,6	151	151	5.6Gbases

Tabla C-1.: Métricas de calidad de las Lecturas raw Illumina reportadas por Velandia-Huerto et. al.(2016) y usadas para el ensamble de novo del organismo *D. vexillum*

D. Pre-tratamientos a las lecturas cortas de Illumina del ADNg de *D. vexillum*

D.1. Control de Calidad Lecturas Cortas Illumina de ADNg de *Didemnum Vexillum* Run1, Sentido Codificante

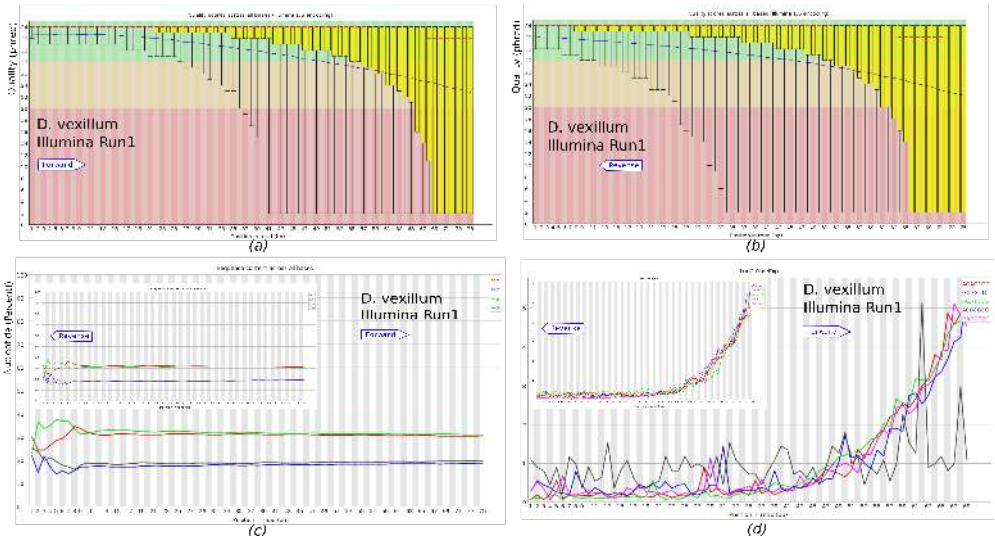


Figura D-1.: Resumen de Evaluación de calidad con el software FastQC para las lecturas raw de illumina del organismo *Didemnum vexillum*: (a),(b) Valor de calidad VS posición de cada nucléotido en la lectura,(c) Porcentaje de Nucléotido en la lectura VS posición de cada nucleotido en la lectura y (d) Log(2) de la Frecuencia de K-mers VS posición de cada nucleotido en la lectura

La figura **D-2** muestra las advertencia producto del analisis de control de calidad con FastQC a las lectura Run1 de Illumina para el Organismo D. vexillum

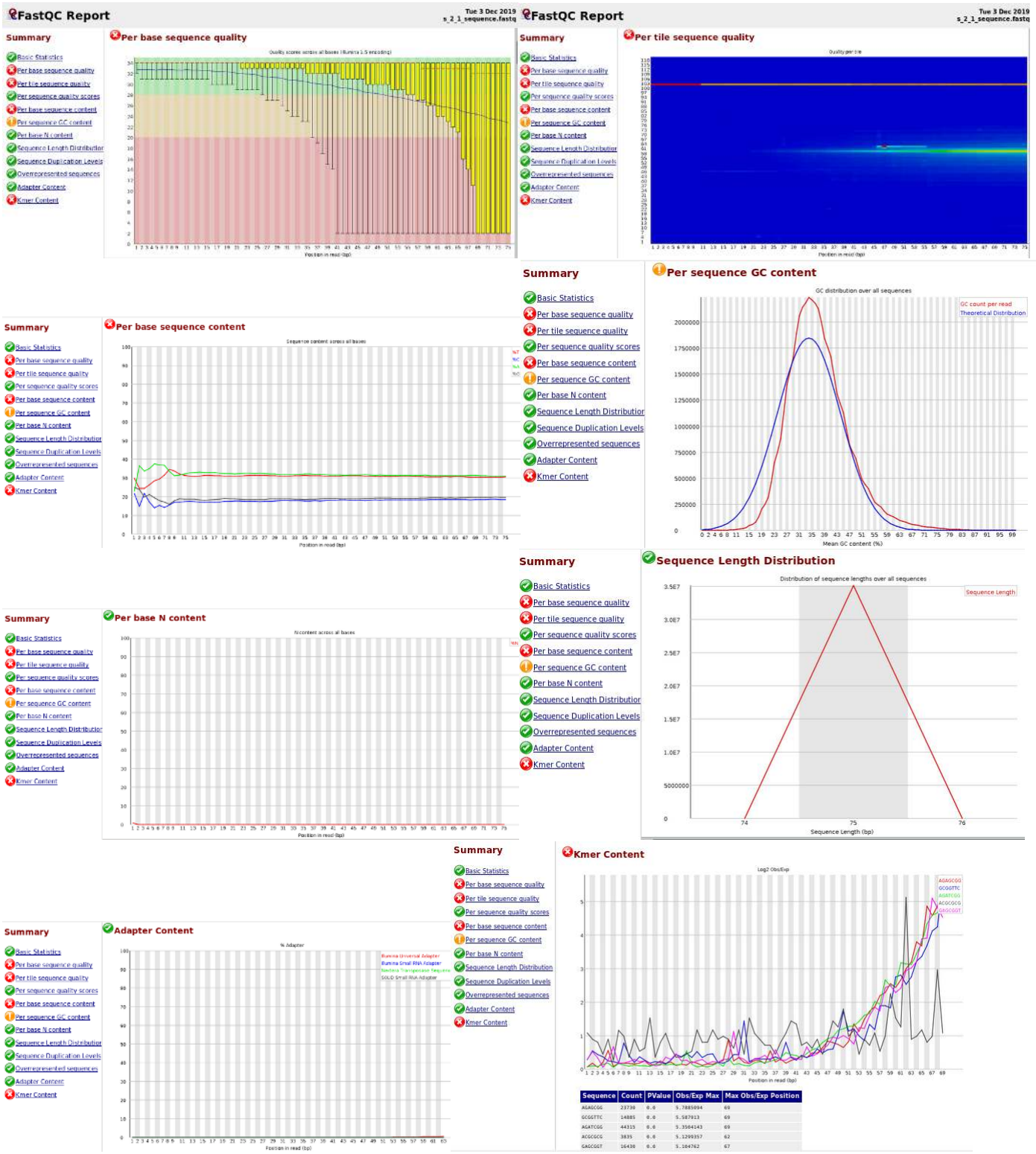


Figura D-2.: Analisis de calidad con FastQC: “Per base sequence quality”, “Per tile sequence quality”,“Per base sequence content”,“Per sequence GC content”,“Sequence Length Distribution”,“Adapter Content”,Kmer Content“ a las lecturas Illumiina Run1, final pareado sentido Forward

D.1.1. Control de Calidad Lecturas Cortas Illumina de ADN de *Didemnum Vexillum* Run1, Antisentido

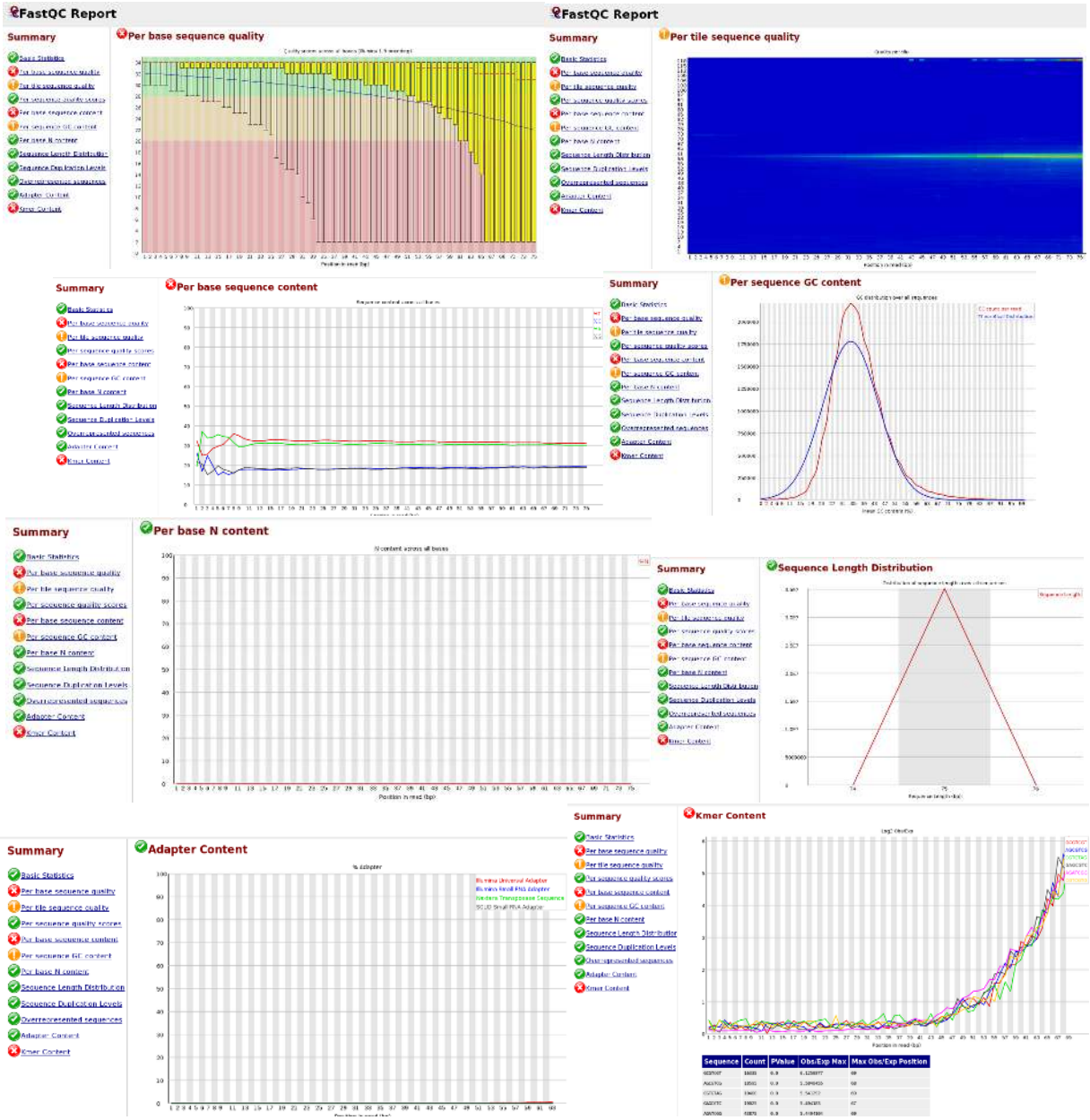


Figura D-3.: Analisis de calidad con FastQC: “Per base sequence quality”, “Per tile sequence quality”,“Per base sequence content”,“Per sequence GC content”,“Sequence Length Distribution”,“Adapter Content”,Kmer Content“ a las lecturas Illumiina Run1, final pareado sentido Reverse

E. Procedimientos exploratorios a las Lecturas Illumina del ADNg de *D. Vexillum*

	Velandia et. al. Raw Illumina	Corte Adaptadores Identificados	Corte de Adaptadores y lecturas phRed <= 10	Corte de Adaptadores y lecturas phRed <= 30
Illumina Run1, Sentido	2.6Gbases	2.3Gbases	1.9Gbases	1.8Gbases
Illumina Run1, Antisentido	2.6Gbases	2.3Gbases	1.9Gbases	1.8Gbases
Illumina Run3, Sentido	5.6Gbases	5.2Gbases	3.3Gbases	3.1Gbases
Illumina Run3, Antisentido	5.6Gbases	5.2Gbases	3.3Gbases	2.8Gbases
Profundidad Aprox.	$30x$	$27x$	$21x$	$17x$

Tabla E-1.: Resultado comparativo del la profundidad de secuenciamiento para un genoma de 550Mbases en D. vexillum. Los datos son tomados desde las lecturas raw de illumina a las que se han aplicado un enmascaramiento y/o recorte de adaptadores.

F. Procedimiento general de Scaffolding

1.Generacion de Concensos CSS:el comando *ccs* toma las sublecturas que comparten un adaptador SMRTbell de PacBio y los combina, luego emplea un modelo estadístico para producir una secuencia consenso de alta calidad por cada plantilla de ADN ciclada y secuenciada como CCS. Figura tomada de `https://github.com/PacificBiosciences/ccs`

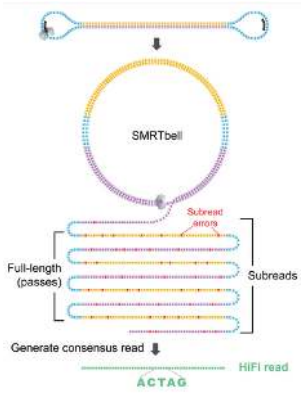


Figura F-1.: Unanimity o Identificación de Concensos Circulares(Circular Consensus Calling)

1. Se descargo e instalo pitchfork: Una coleccion de scritps que crean software para Pacbio a partir de proyectos hospedados en "gitHub" https://github.com/PacificBiosciences/pitchfork
2. El programa necesita como entrada una archivo subreads.bam No alineado que contiene las sublecturas de cada SMRTbell secuecnciada. Esto fue hecho con la herramienta bax2bam instalada en el paso 1. 3. se ejecuta en la linea de commandos
4. Se obtuvieron las lecturas CSS en formato FASTA

G. Identificación y Enmascaramiento de regiones repetitivas

Además de las secuencias que codifican para proteínas funcionales y elementos de regulación genética, la mayoría de los genomas de eucariotas tambíen poseen un gran número de secuencias no-codificantes que se repiten masivamente en todo el genoma [Berman *et al.*, 2004]. Aunque la presencia de repeticiones por su número y tamaño pueden consumir recursos en las tareas de biología computacional o, la búsqueda de homologia que incluya estas secuencias contra bases de datos que mantiene secuencias que codifican proteinas puede dar lugar a falsas interpretaciones de los resultados [Tørresen *et al.*, 2019], una identificación y anotación de las regiones repetitivas del genoma tambíen busca no aislarlas de su contexto biologico, debido a que aún no es claro como pueden afectar la funcion del ADN [Sokol *et al.*, 2007]. La identificación de secuencias repetitivas en el re-ensamble genómico de *D. vexillum* usó métodos de predicciones ab initio y comparacion con secuencias homólogas en C. intestinalis, el procedimiento general se puede visual en la figura **G-1**.

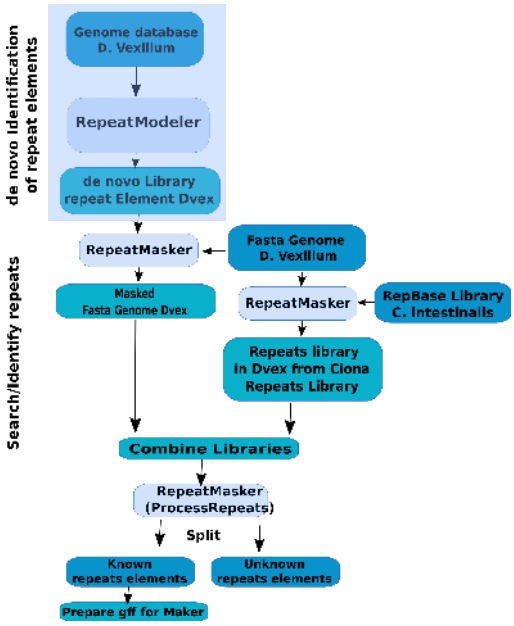


Figura G-1.: Procedimiento general para la identificacion de secuencias repetitivas en el re-ensamble genómico de D. vexillum

La identificación de secuencias repetitivas *de novo* uso el software RepeatModeler v.1.0.4 [Smit y Hubley, 2019] con los parámetros por omision y usando los scaffolds del ensamble genómico de *D.vexillum*. El enmascaramiento de las repeticiones y las secuencias de baja complejidad sobre el ensamble se realizó con RepeatMasker v.open-4.0.5 [Smit *et al.*, 2015] usando la libreria de repeticiones de novo en combinación con la librerias de repeticiones de C. intestinalis de RepBase Version 20.03 [Bao *et al.*, 2015].
Esta etapa finaliza con la identificación de 293Mbases(56.48 %) en 95627(87.1 %) scaffolds de elementos repetitivos clasificados que, conformaron la librería de repeticiones usada en los procesos posteriores de anotación genomica. Los elementos repetitivos identificados pero no clasificados fueron excluidos de la anotación genómica del organismo *D. vexillum*.

Bibliografía

[Abouelhoda *et al.*, 2002] Abouelhoda, M. I., Kurtz, S., y Ohlebusch, E. (2002). The enhanced suffix array and its applications to genome analysis. En *International Workshop on Algorithms in Bioinformatics*, pp. 449–463. Springer.

[Afgan *et al.*, 2018] Afgan, Enis and Baker, Dannon and Batut, Bérénice and Van Den Beek, Marius and Bouvier, Dave and Čech, Martin and Chilton, John and Clements, Dave and Coraor, Nate and Grüning, Björn A and others (2018). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46(W1):W537–W544.

[Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., y Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

[2010] Andrews, Simon and others (2010). Fastqc: a quality control tool for high throughput sequence data.

[aniseed, 2020] aniseed (2020). species. `https://www.aniseed.cnrs.fr/aniseed/species`. Online; accessed 22-Enero-2020.

[Bao *et al.*, 2015] Bao, W., Kojima, K. K., y Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna*, 6(1):11.

[Berman *et al.*, 2004] Berman, P., Bertone, P., Dasgupta, B., Gerstein, M., Kao, M.-Y., y Snyder, M. (2004). Fast optimal genome tiling with applications to microarray design and homology search. *J Comput Biol*, 11(4):766–85.

[Berná y Alvarez-Valin, 2014] Berná, L. y Alvarez-Valin, F. (2014). Evolutionary genomics of fast evolving tunicates. *Genome biology and evolution*, 6(7):1724–1738.

[Biosciences, 2014] Biosciences, P. (2014). Post run qc analisis. `https://www.pacb.com/training/PostRunQCAnalysis/story_content/external_files/Post%20Run%20QC%20Analysis.pdf`. Online; accessed 7-December-2019.

[Biosciences, 2015a] Biosciences, P. (2015a). Procedure and checklist 10 kb template preparation and sequencing. `https://www.pacb.com/wp-content/uploads/2015/09/Procedure-Checklist-10-kb-Template-Preparation-and-Sequencing.pdf`. Online; accessed 7-December-2019.

[Biosciences, 2015b] Biosciences, P. (2015b). Procedure and checklist 20 kb template preparation using bluepippintm size-selection system. https://www.pacb.com/wp-content/uploads/2015/09/ProcedureChecklist-20-kb-Template-Preparation-Using-BluePippin-Size-Selection.pdf. [Online; accessed 7-December-2019.

[Biosciences, 2015c] Biosciences, P. (2015c). Procedure and checklist 5 kb template preparation and sequencing. https://www.pacb.com/wp-content/uploads/2015/09/Procedure-Checklist-5-kb-Template-Preparation-and-Sequencing.pdf. [Online; accessed 7-December-2019.

[Birney *et al.*, 2004] E. Birney and D. Andrews and P. Bevan and M. Caccamo and G. Cameron and Y. Chen and L. Clarke and G. Coates and T. Cox and J. Cuff and others (2004). Ensembl 2004. *Nucleic acids research*, 32(suppl 1):D468.

[Blanchoud *et al.*, 2018a] Blanchoud, S., Rutherford, K., Zondag, L., Gemmell, N. J., y Wilson, M. J. (2018a). De novo draft assembly of the botrylloides leachii genome provides further insight into tunicate evolution. *Scientific Reports*, 8(1):5518.

[Blanchoud *et al.*, 2018b] Blanchoud, S., Rutherford, K., Zondag, L., Gemmell, N. J., y Wilson, M. J. (2018b). De novo draft assembly of the Botrylloides leachii genome provides further insight into tunicate evolution. *Sci Rep*, 8(1):5518.

[Boetzer y Pirovano, 2014] Boetzer, M. y Pirovano, W. (2014). Sspace-longread: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics*, 15(1):211.

[Brozovic *et al.*, 2018] Brozovic, Matija and Dantec, Christelle and Dardaillon, Justine and Dauga, Delphine and Faure, Emmanuel and Gineste, Mathieu and Louis, Alexandra and Naville, Magali and Nitta, Kazuhiro R and Piette, Jacques and others (2018). Aniseed 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. *Nucleic acids research*, 46(D1):D718–D725.

[Bushnell, 2016] Bushnell, B. (2016). Bbmap. https://sourceforge.net/projects/bbmap. Online; accessed 1-December-2017.

[Bushnell *et al.*, 2017] Bushnell, B., Rood, J., y Singer, E. (2017). Bbmerge–accurate paired shotgun read merging via overlap. *PLoS One*, 12(10):e0185056.

[Campbell *et al.*, 2014] Campbell, M. S., Holt, C., Moore, B., y Yandell, M. (2014). Genome annotation and curation using maker and maker-p. *Current Protocols in Bioinformatics*, 48(1):4–11.

[Cantarel *et al.*, 2008] Cantarel, B., Korf, I., Robb, S., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., y Yandell, M. (2008). Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, 18(1):188.

[Consortium, 2006] Consortium, G. O. (2006). The gene ontology (go) project in 2006. *Nucleic acids research*, 34(suppl_1):D322–D326.

[Contreras *et al.*, 2016] Contreras, A. V., Cocom-Chan, B., Hernandez-Montes, G., Portillo-Bobadilla, T., y Resendis-Antonio, O. (2016). Host-microbiome interaction and cancer: potential application in precision medicine. *Frontiers in physiology*, 7:606.

[Crabtree *et al.*, 2007] Crabtree, J., Angiuoli, S. V., Wortman, J. R., y White, O. R. (2007). Sybil: methods and software for multiple genome comparison and visualization. En *Gene Function Analysis*, pp. 93–108. Springer.

[Darren *et al.*, 2004] Darren, C., Dave, C., y Andy, W. (2004). Perl template toolkit.

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

[Delsuc *et al.*, 2018] Delsuc, F., Philippe, H., Tsagkogeorga, G., Simion, P., Tilak, M. K., Turon, X., Lopez-Legentil, S., Piette, J., Lemaire, P., y Douzery, E. J. P. (2018). A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biol.*, 16(1):39.

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

[Drillon *et al.*, 2014a] Drillon, G., Carbone, A., y Fischer, G. (2014a). SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS ONE*, 9(3):e92621.

[Drillon *et al.*, 2014b] Drillon, G., Carbone, A., y Fischer, G. (2014b). Synchro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One*, 9(3).

[Eid *et al.*, 2009] Eid, John and Fehr, Adrian and Gray, Jeremy and Luong, Khai and Lyle, John and Otto, Geoff and Peluso, Paul and Rank, David and Baybayan, Primo and Bettman, Brad and others (2009). Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

[Eilbeck *et al.*, 2005] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., y Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome Biol*, 6(5):R44.

[Förster *et al.*, 2014] Förster, F., Schultz, J., Hedrich, R., y Hackl, T. (2014). proovread : large-scale high-accuracy pacbio correction through iterative short read consensus. *Bioinformatics*, 30(21):3004–3011.

[Gassmann y McHoull, 2014] Gassmann, M. y McHoull, B. (2014). Dna integrity number (din) with the agilent 2200 tapestation system and the agilent genomic dna screentape assay. *Agilent Technologies*.

[Grabherr *et al.*, 2011] Grabherr, Manfred G and Haas, Brian J and Yassour, Moran and Levin, Joshua Z and Thompson, Dawn A and Amit, Ido and Adiconis, Xian and Fan, Lin and Raychowdhury, Raktima and Zeng, Qiandong and others (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644.

[Grabherr *et al.*, 2010] Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alf?ldi, J., Di Palma, F., y Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, 26(9):1145–1151.

[Grabherr, 2010] Grabherr, Manfred G, P. M. M. M. E. A. J. D. P. F. L.-T. K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, 26(9):1145–1151.

[Haas *et al.*, 2016] Haas, B and Papanicolaou, A and others (2016). Transdecoder (find coding regions within transcripts). *Google Scholar*.

[Haas *et al.*, 2004] Haas, B. J., Delcher, A. L., Wortman, J. R., y Salzberg, S. L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18):3643–3646.

[Henneke y Golenbock, 2004] Henneke, P. y Golenbock, D. T. (2004). Phagocytosis, innate immunity, and host–pathogen specificity. *The Journal of experimental medicine*, 199(1):1–4.

[Hill *et al.*, 2008] Hill, M. M., Broman, K. W., Stupka, E., Smith, W. C., Jiang, D., y Sidow, A. (2008). The c. savignyi genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Research*, 18(8):1369–1379.

[Hirose, 2009] Hirose, E. (2009). Ascidian tunic cells: morphology and functional diversity of free cells outside the epidermis. *Invertebrate Biology*, 128(1):83–96.

[Hu *et al.*, 2008] Hu, Z.-L., Bao, J., y Reecy, J. M. (2008). Categorizer: a web-based program to batch analyze gene on-tology classification categories. *Online J Bioinform*, 9:108–112.

[Illumina, 2017] Illumina (2017). Phix control library. `https://support.illumina.com/bulletins/2017/02/what-is-the-phix-control-v3-library-and-what-is-its-function-in-.html`. Online; accessed 1-December-2017.

[Jue *et al.*, 2016a] Jue, N. K., Batta-Lona, P. G., Trusiak, S., Obergfell, C., Bucklin, A., O'Neill, M. J., y O'Neill, R. J. (2016a). Rapid Evolutionary Rates and Unique Genomic Signatures Discovered in the First Reference Genome for the Southern Ocean Salp, Salpa thompsoni (Urochordata, Thaliacea). *Genome Biol Evol*, 8(10):3171–3186.

[Jue *et al.*, 2016b] Jue, N. K., Batta-Lona, P. G., Trusiak, S., Obergfell, C., Bucklin, A., O'Neill, M. J., y O'Neill, R. J. (2016b). Rapid evolutionary rates and unique genomic signatures discovered in the first reference genome for the southern ocean salp, salpa thompsoni (urochordata, thaliacea). *Genome Biology and Evolution*, 8(10):3171–3186.

[Karp *et al.*, 2010] Karp, Peter D and Paley, Suzanne M and Krummenacker, Markus and Latendresse, Mario and Dale, Joseph M and Lee, Thomas J and Kaipa, Pallavi and Gilham, Fred and Spaulding, Aaron and Popescu, Liviu and others (2010). Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 11(1):40–79.

[Kasprzyk, 2011] Kasprzyk, A. (2011). Biomart: driving a paradigm change in biological data management. *Database*, 2011.

[Lawson *et al.*, 2006] D. Lawson and P. Arensburger and P. Atkinson and N.J. Besansky and R.V. Bruggner and R. Butler and K.S. Campbell and G.K. Christophides and S. Christley and E. Dialynas and others (2006). Vectorbase: a home for invertebrate vectors of human pathogens. *Nucleic Acids Research*, 35(suppl 1):D503.

[Lee *et al.*, 2009] Lee, E., Harris, N., Gibson, M., Chetty, R., y Lewis, S. (2009). Apollo: a community resource for genome annotation editing. *Bioinformatics*, 25(14):1836–1837.

[Little y Seehaus, 1988] Little, M. y Seehaus, T. (1988). Comparative analysis of tubulin sequences. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 90(4):655–670.

[Liu *et al.*, 2018] Liu, D., Hunt, M., y Tsai, I. J. (2018). Inferring synteny between genome assemblies: a systematic evaluation. *BMC Bioinformatics*, 19(1):26.

[Mardis, 2013] Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, 6(1):287–303. PMID: 23560931.

[Meyer *et al.*, 2009] Meyer, M., Munzner, T., y Pfister, H. (2009). Mizbee: a multiscale synteny browser. *IEEE transactions on visualization and computer graphics*, 15(6):897–904.

[Mungall *et al.*, 2007] Mungall, C. J., Emmert, D. B., y Consortium, F. (2007). A chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23(13):i337–i346.

[Myers, 2016] Myers, E. (2016). Daligner. `https://github.com/thegenemyers/DALIGNER`. [Online; accessed 1-December-2017.

[Myers *et al.*, 2000] Myers, Eugene W and Sutton, Granger G and Delcher, Art L and Dew, Ian M and Fasulo, Dan P and Flanigan, Michael J and Kravitz, Saul A and Mobarry, Clark M and Reinert, Knut HJ and Remington, Karin A and others (2000). A whole-genome assembly of drosophila. *Science*, 287(5461):2196–2204.

[Myers, 2015] Myers, G. (2015). Dextractor. `https://github.com/thegenemyers/DEXTRACTOR`. Online; accessed 1-December-2017.

[NIGMS, 2002] NIGMS, N. (2002). Developing robust components for model organism databases. `https://grants.nih.gov/grants/guide/rfa-files/rfa-hg-02-002.html`. [Online; accessed 1-May-2020.

[O'Connor *et al.*, 2008] O'Connor, B. D., Day, A., Cain, S., Arnaiz, O., Sperling, L., y Stein, L. D. (2008). Gmodweb: a web framework for the generic model organism database. *Genome Biol*, 9(6):R102.

[Ordóñez *et al.*, 2015] Ordóñez, V., Pascual, M., Fernández-Tejedor, M., Pineda, M., Tagliapietra, D., y Turon, X. (2015). Ongoing expansion of the worldwide invader didemnum vexillum (ascidiacea) in the mediterranean sea: high plasticity of its biological cycle promotes establishment in warm waters. *Biological invasions*, 17(7):2075–2085.

[PacBio, 2015] PacBio (2015). Unanimity. `https://github.com/PacificBiosciences/ccs`. [Online; accessed 1-July-2019.

[Proost *et al.*, 2012] Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., y Vandepoele, K. (2012). i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.*, 40(2):e11.

[Pryszcz y Gabaldón, 2016] Pryszcz, L. P. y Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, 44(12):e113–e113.

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

[Qin *et al.*, 2019] Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L., y Ruan, J. (2019). Lrscaf: improving draft genomes using long noisy reads. *BMC genomics*, 20(1):955.

[Quevillon *et al.*, 2005] Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., y Lopez, R. (2005). Interproscan: protein domains identifier. *Nucleic acids research*, 33(suppl_2):W116–W120.

[Reeves *et al.*, 2008] Reeves, G. A., Eilbeck, K., Magrane, M., O'Donovan, C., Montecchi-Palazzi, L., Harris, M. A., Orchard, S., Jimenez, R. C., Prlic, A., Hubbard, T. J. P., Hermjakob, H., y Thornton, J. M. (2008). The protein feature ontology: a tool for the unification of protein feature annotations. *Bioinformatics*, 24(23):2767–72.

[Salmela y Rivals, 2014] Salmela, L. y Rivals, E. (2014). Lordec: accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514.

[Sanderson *et al.*, 2013] Sanderson, L.-A., Ficklin, S. P., Cheng, C.-H., Jung, S., Feltus, F. A., Bett, K. E., y Main, D. (2013). Tripal v1. 1: a standards-based toolkit for construction of online genetic and genomic databases. *Database*, 2013.

[Satou *et al.*, 2019] Satou, Y., Nakamura, R., Yu, D., Yoshida, R., Hamada, M., Fujie, M., Hisata, K., Takeda, H., y Satoh, N. (2019). A nearly complete genome of ciona intestinalis type a (c. robusta) reveals the contribution of inversion to chromosomal evolution in the genus ciona. *Genome biology and evolution*, 11(11):3144–3157.

[Sensen, 2005] Sensen, C. W. (2005). *Handbook of genome research: genomics, proteomics, metabolomics, bioinformatics, ethics and legal issues; Vol. 1 und 2*. Wiley-VCH.

[Seo *et al.*, 2001] Seo, Hee-Chan and Kube, Michael and Edvardsen, Rolf B and Jensen, Marit F and Beck, Alfred and Spriet, Endy and Gorsky, Gabriel and Thompson, Eric M and Lehrach, Hans and Reinhardt, Richard and others (2001). Miniature genome in the marine chordate oikopleura dioica. *Science*, 294(5551):2506–2506.

[Shuhua *et al.*, 2019] Shuhua, F., Anqi, W., y Fai, A. K. (2019). A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biology*, 20(1):26.

[Simillion *et al.*, 2008] Simillion, C., Janssens, K., Sterck, L., y Van de Peer, Y. (2008). i-adhore 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, 24(1):127–128.

[Simão *et al.*, 2015] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., y Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

[Skelly, 2014] Skelly, T. (2014). Pacbioeda. `https://github.com/TomSkelly/PacBioEDA`. Online; accessed 4-December-2019.

[Skinner *et al.*, 2009] Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., y Holmes, I. H. (2009). Jbrowse: a next-generation genome browser. *Genome research*, 19(9):1630–1638.

[Slater y Birney, 2005] Slater, G. S. C. y Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31.

[Small *et al.*, 2007] Small, K. S., Brudno, M., Hill, M. M., y Sidow, A. (2007). A haplome alignment and reference sequence of the highly polymorphic ciona savignyi genome. *Genome biology*, 8(3):R41.

[Smit y Hubley, 2019] Smit, A. y Hubley, R. (2019). Repeatmodeler-1.0. 11. *Institute for Sys-tems Biology. http://www. repeatmasker. org/RepeatModeler/. Accessed*, 15.

[Smit *et al.*, 2015] Smit, A., Hubley, R., y Green, P. (2015). Repeatmasker open-4.0. 2013–2015.

[Smith *et al.*, 2012] Smith, Richard N and Aleksic, Jelena and Butano, Daniela and Carr, Adrian and Contrino, Sergio and Hu, Fengyuan and Lyne, Mike and Lyne, Rachel and Kalderimis, Alex and Rutherford, Kim and others (2012). Intermine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, 28(23):3163–3165.

[Sokol *et al.*, 2007] Sokol, D., Benson, G., y Tojeira, J. (2007). Tandem repeats over the edit distance. *Bioinformatics*, 23(2):e30–e35.

[Stein, 2003] Stein, L. D. (2003). Integrating biological databases. *Nature Reviews Genetics*, 4(5):337–345.

[Stein *et al.*, 2002] Stein, Lincoln D and Mungall, Christopher and Shu, ShengQiang and Caudy, Michael and Mangone, Marco and Day, Allen and Nickerson, Elizabeth and Stajich, Jason E and Harris, Todd W and Arva, Adrian and others (2002). The generic genome browser: a building block for a model organism system database. *Genome research*, 12(10):1599–1610.

[Stuart y Ezekowitz, 2008] Stuart, L. M. y Ezekowitz, R. A. (2008). Phagocytosis and comparative innate immunity: learning on the fly. *Nature Reviews Immunology*, 8(2):131–141.

[Tedersoo *et al.*, 2018] Tedersoo, L., Tooming-Klunderud, A., y Anslan, S. (2018). Pacbio metabarcoding of fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist*, 217(3):1370–1385.

[Tischler y Myers, 2017] Tischler, G. y Myers, E. W. (2017). Non hybrid long read consensus using local de bruijn graph assembly. *bioRxiv*, p. 106252.

[Tørresen *et al.*, 2019] Tørresen, Ole K and Star, Bastiaan and Mier, Pablo and Andrade-Navarro, Miguel A and Bateman, Alex and Jarnot, Patryk and Gruca, Aleksandra and Grynberg, Marcin and Kajava, Andrey V and Promponas, Vasilis J and others (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic acids research*, 47(21):10994–11006.

[Valentine *et al.*, 2006] Valentine, P. C., Carman, M. R., Blackwood, D. S., y Heffron, E. J. (2006). Ecological observations on the colonial ascidian didemnum sp. in a new england tide pool habitat.

[Velandia *et al.*, 2016] Velandia, H. C., Gittenberger, A. A., Brown, F. D., Stadler, P. F., y Bermúdez-Santana, C. I. (2016). Automated detection of ncrnas in the draft genome sequence of a colonial tunicate: the carpet sea squirt didemnum vexillum. *BMC Genomics*, 17:691.

[Venter *et al.*, 2001] Venter, J Craig and Adams, Mark D and Myers, Eugene W and Li, Peter W and Mural, Richard J and Sutton, Granger G and Smith, Hamilton O and Yandell, Mark and Evans, Cheryl A and Holt, Robert A and others (2001). The sequence of the human genome. *science*, 291(5507):1304–1351.

[Voskoboynik *et al.*, 2013] Voskoboynik, A., Neff, N. F., Sahoo, D., Newman, A. M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H. C., Mantalas, G. L., Palmeri, K. J., Ishizuka, K. J., Gissi, C., Griggio, F., Ben-Shlomo, R., Corey, D. M., Penland, L., White 3rd, R. A., Weissman, I. L., y Quake, S. R. (2013). The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife*, 2:e00569.

[Wang *et al.*, 2019] Wang, S., Harting, J., Tseng, E., y Baybayan, P. (2019). Getting the most out of your pacbio® libraries with size selection.

[Wang *et al.*, 2012] Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T. H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., y Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, 40(7):e49.

[Wu *et al.*, 2016] Wu, T. D., Reeder, J., Lawrence, M., Becker, G., y Brauer, M. J. (2016). Gmap and gsnap for genomic sequence alignment: enhancements to speed, accuracy, and functionality. En *Statistical Genomics*, pp. 283–334. Springer.

[Youens-Clark *et al.*, 2009] Youens-Clark, K., Faga, B., Yap, I. V., Stein, L., y Ware, D. (2009). Cmap 1.01: a comparative mapping application for the internet. *Bioinformatics*, 25(22):3040–3042.

[Zhao y Schranz, 2019] Zhao, T. y Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proceedings of the National Academy of Sciences*, 116(6):2165–2174.