FACULTAD DE CIENCIAS AGRARIAS

UNIVERSIDAD NACIONAL DE ROSARIO

USO DE DATOS DE TRANSCRIPTÓMICA PARA LA ANOTACIÓN DE GENES EN EL GENOMA DE PACÚ (*Piaractus mesopotamicus*)

Lic. Victoria María Posner

TRABAJO FINAL PARA OPTAR AL TITULO DE ESPECIALISTA EN BIOINFORMÁTICA

DIRECTORA: Dra. Florencia C. Mascali

AÑO 2022

USO DE DATOS DE TRANSCRIPTÓMICA PARA LA ANOTACIÓN DE GENES EN EL GENOMA DE PACÚ (*Piaractus mesopotamicus*)

Victoria María Posner

Licenciada en Biotecnología – Universidad Nacional de Rosario

Este Trabajo Final es presentado como parte de los requisitos para optar al grado académico de Especialista en Bioinformática, de la Universidad Nacional de Rosario y no ha sido previamente presentada para la obtención de otro título en esta u otra Universidad. El mismo contiene los resultados obtenidos en investigaciones llevadas a cabo en el Laboratorio Mixto de Biotecnología Acuática (UNR-SCTeI Santa Fe), durante el período comprendido entre abril del 2021 y octubre del 2022, bajo la dirección de la Dra Florencia C. Mascali.

Nombre y firma del autor

Nombre y firma del Director

Defendida: .....de 20__.

ABREVIATURAS Y SÍMBOLOS

- **RGA:** Recursos genéticos acuáticos.
- **NGS:** *Next Generation Sequencing***.** Secuenciacion de nueva generación.
- **UTRs:** *Untranslated región.* Región no traducida.
- **EST:** *Expressed Sequence Tag.* Marcador de secuencia expresada.
- **CDS:** *Coding sequence*. Región codificante.
- **ORF:** *Open Reading Frame.* Marco de lectura abierto.
- **pb:** pares de bases.
- **AED:** *Annotation Edit Distance.* Parámetro que mide la bondad de ajuste de una anotación a la evidencia que la respalda.
- **ENA:** *European Nucleotie Archive.* Archivo europeo de nucleótidos

RESUMEN

En Argentina, la producción de pacú (*Piaractus mesopotamicus*) ocupa el primer lugar dentro de las producciones por piscicultura de especies nativas. A pesar de la importancia de la misma, existe poca información genómica sobre esta especie, lo cual limita el desarrollo de tecnología y de programas de mejoramiento genético. En el Laboratorio Mixto de Biotecnología Acuática fueron ensamblados recientemente un genoma de referencia para hembra y uno para macho a partir de la tecnología de secuenciación corta de Illumina. Sin embargo, los mismos aún no han sido anotados. Profundizar el análisis de dichos genomas permitirá no solo aumentar el conocimiento sobre diferentes aspectos de la biología del pacú, sino que también dará lugar al diseño y desarrollo de nuevos proyectos. Es por ello que el objetivo de este trabajo fue contribuir a la anotación de dichos genomas utilizando datos de transcriptómica disponibles para esta especie.

En el presente trabajo se utilizaron secuencias de un transcriptoma de músculo de pacú provenientes de un repositorio público. Se realizó un análisis de calidad de las secuencias crudas y luego se ensamblaron de manera conjunta usando el programa Trinity. El análisis del ensamblado indicó que el mismo es de alta calidad, con un alto aprovechamiento de las secuencias crudas (98.84 %) y alta presencia (81.4%) de ortólogos de ciertos genes que son universales, se expresan persistentemente y se presentan casi exclusivamente como copias únicas en el genoma. Para la anotación funcional se combinaron datos de predicción de secuencias proteicas, análisis de homologías de secuencias y predicción de dominios, de péptidos señales y de familias. Al anotar el transcriptoma se encontraron 64.111 genes e isoformas, indicando una perspectiva de alta utilidad en el uso de estas secuencias para anotar los genomas.

Los genomas obtenidos para un macho y una hembra de pacú a partir de tecnología Illumina se encuentran altamente fragmentados. Esto trajo como consecuencia una gran dificultad computacional en el proceso de anotación de los mismos, a pesar de trabajar en un servidor con un alto poder de cómputo. Por esto, solo se llevó a cabo una prueba de anotación de un fragmento de cada genoma. El genoma de la hembra se procesó en un 38%, anotando 5.970 genes, con un largo promedio de 4.660,44 bases. El genoma del macho, en cambio, se procesó en un 23%, anotando 4.182 genes, con un largo promedio de 3.953,42 bases A pesar de ser resultados preliminares los mismos son alentadores, ya que presentan altos niveles de calidad según el parámetro AED.

Palabras clave: Transcriptoma, Genoma, Anotación.

ABSTRACT

In Argentina, the production of pacú (*Piaractus mesopotamicus*) occupies the first place within the productions by fish farming of native species. Despite its importance, there is little genomic information on this species, which limits the development of technology and breeding programs. A female and a male reference genome were recently assembled at the Joint Aquatic Biotechnology Laboratory using Illumina short-sequencing technology. However, they have not yet been annotated. Deepening the analysis of these genomes will allow not only to increase knowledge about different aspects of the biology of the pacú, but will also lead to the design and development of new projects. That is why the aim of this work was to contribute to

In the present work, sequences of a pacú muscle transcriptome from a public repository were used. Quality analysis of the raw sequences was performed and these were then assembled together using the Trinity program. The analysis of the assembly indicated that it is of high quality, with a high use of raw sequences (98.84%) and a high presence (81.4%) of orthologs of certain genes that are universal, are persistently expressed and are presented almost exclusively as unique copies in the genome. Data from protein sequence prediction, sequence homology analysis, and domain, signal peptide, and family prediction were combined for functional annotation. Annotating the transcriptome found 64,111 genes and isoforms, indicating a prospect of high utility in using these sequences to annotate genomes.

Genomes obtained for one male and one female pacú from Illumina technology are highly fragmented. This resulted in great computational difficulty in the annotation process, despite working on a server with high computing power. For this reason, only one annotation test of a fragment of each genome was carried out. The female genome was 38% processed, scoring 5,970 genes, with an average length of 4,660.44 bases. The male genome, on the other hand, was processed in 23%, annotating 4,182 genes, with an average length of 3,953.42 bases. Despite being preliminary results, they are encouraging, since they present high levels of quality according to the AED parameter.

Key words: Transcriptome, Genome, Annotation

ÍNDICE

INTRODUCCIÓN

El pacú, *Piaractus mesopotamicus,* es un pez de agua dulce perteneciente a Sudamérica que se distribuye en la Cuenca del Plata, en los ríos Paraná, Paraguay y Uruguay, cuya mayor distribución se produce en las planicies inundadas del Pantanal (Liotta, 2005; Resende, 2003). Es una de las especies no modelo más importante cultivada en América del Sur en los países de Colombia, Perú, Venezuela, Argentina y Brasil; y su producción ha aumentado incluso en otras partes del mundo, como China, Myanmar, Tailandia y Vietnam. (Flores Nava, 2007; Honglang, 2007; Laurenti et al., 2014). Argentina, la producción de pacú ocupa el primer lugar dentro de las producciones por piscicultura de especies nativas. El cultivo se desarrolla principalmente de manera extensiva en tanques excavados en tierra en las provincias de Misiones, Chaco, Formosa y Corrientes. Existen también experiencias de cultivo en las provincias de Santa Fe y Entre Ríos, aunque en menor magnitud debido a que las temperaturas menos cálidas limitan los cultivos. Es una especie con mercado instalado, de gran aceptación a nivel nacional e internacional. Actualmente, los niveles de producción alcanzados no llegan a cubrir la demanda existente.

El desarrollo de una acuicultura sustentable desde un punto de vista ambiental y económico requiere la comprensión de la base genética de los rasgos que limitan y/o mejoran el desarrollo del cultivo. Las condiciones de producción y de mercado indican que un plan de mejoramiento genético para pacú podría tener un impacto positivo en la producción. El desarrollo de biotecnología y selección de genotipos superiores para pacú podría maximizar la producción sustentable de este pez, principalmente si herramientas genómicas fueran desarrolladas para su aplicación de rutina en programas de manejo, producción y mejoramiento genético. Sin embargo, la mayoría de los productores de pacú de la región, no tienen un programa de manejo de reproductores ni existe un estudio sobre variabilidad genética de poblaciones silvestres y de cautiverio, hecho que amenaza la sostenibilidad de la producción. Tampoco existen líneas seleccionadas de alto crecimiento como en el caso de salmónidos o de tilapia, dos de las especies exóticas de mayor producción en Sudamérica.

Los recursos genéticos acuáticos (RGA) comprenden todo material de naturaleza biológica y acuática que contenga información genética de valor o utilidad real o potencial. Son una dimensión de la Biodiversidad, la cual se estratifica desde genes, hacia individuos, especies, poblaciones, ecosistemas y paisajes. Los recursos genéticos acuáticos se encuentran en la base de la productividad y sostenibilidad de la acuicultura. La Comisión de Recursos Genéticos para la Alimentación y la Agricultura (CGRFA) de la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO), en su 11ava reunión ordinaria (Laurenti et al., 2014), reconoció la importancia y la vulnerabilidad de los RGA. Sin embargo, a pesar de su relevancia, la información sobre RGA que sirva de base para la gestión de los recursos es escasa en la mayoría de los países. En el caso del pacú, por ejemplo, existe poca información

genómica, estando disponibles solamente el ADN mitocondrial (Pimentel et al., 2014), un mapa de ligamiento (Mastrochirico-Filho, Borges, et al., 2020) y transcriptomas de músculo (Mareco et al., 2015), de hígado (Mastrochirico-Filho et al., 2016) y de hígado ante infección con un patógeno (Mastrochirico-Filho, Hata, et al., 2020), lo cual limita el desarrollo de tecnología y de programas de mejoramiento genético. Aumentar la información genómica sobre la especie permitirá el desarrollo de tecnología con un significativo aporte a los programas de mejoramiento genético.

Desde el advenimiento de las tecnologías de secuenciación de nueva generación (NGS) en la década del 2000, la cantidad de datos sobre secuencias biológicas ha aumentado de manera exponencial, en forma concordante con la reducción continua de los costos de secuenciación. Además, dio lugar al desarrollo de nuevas aplicaciones (Dunham et al., 2014), permitiendo la generación de recursos genéticos para un mayor número de especies (Huete-Pérez & Quezada, 2013). Entre ellos, la obtención del genoma para una determinada especie de pez representa el punto de partida a partir del cual es posible avanzar significativamente en el estudio de otras áreas de la acuicultura tales como reproducción, sanidad, nutrición, comportamiento, fisiología, larvicultura y biología molecular, entre otras (Purcell et al., 2018).

La generación de un genoma de referencia involucra varias etapas luego de la obtención de las secuencias crudas, que incluyen el análisis y filtrado de los datos de baja calidad, el ensamblado de las secuencias restantes y su posterior anotación. La etapa de anotación implica la descripción de las diferentes características de las secuencias ensambladas, las cuales pueden ser características estructurales (ontología de secuencia o SO: exones, intrones, UTRs (del inglés *Untranslated region*, por regiones no traducidas), empalmes alternativos, etc.) o funcionales (ontología de genes o GO: describir en qué proceso están involucrados, su función molecular, localización de la expresión, etc.). Aunque la secuenciación de especies no modelo se ha vuelto más accesible, la anotación del genoma representa un gran desafío, principalmente en especies eucariotas con genomas de gran tamaño. Varios factores son responsables de esto. En primer lugar, las longitudes de lectura cortas de las plataformas de secuenciación de segunda generación, como Illumina, implican que sus ensamblados tienen baja contigüidad. En segundo lugar, la naturaleza exótica de muchos de los genomas secuenciados dificulta la búsqueda de genes al no existir secuencias de referencia para su anotación. Esto dificulta entrenar, optimizar y configurar herramientas de predicción y anotación de genes (Yandell & Ence, 2012). En tercer lugar, la multitud de fuentes de datos de secuenciación y las diferentes aproximaciones que deben seguirse según el tipo de organismo agregan aún más complejidad a la tarea y hace necesario un alto grado de conocimiento de algoritmos y protocolos de procesamiento de los datos. Además, son necesarios tiempos y recursos computacionales considerables, ya que algunas herramientas de ensamblado o anotación pueden llevar varias semanas (Dominguez Del Angel et al., 2018). Para sortear estas dificultades, una de las mejores evidencias termina siendo los datos detallados de EST (del inglés *expresed sequence tag*, por marcador de

secuencia expresada) o RNA-seq, ya que proporcionan modelos de genes con información sobre sitios de corte y empalme, sitios de inicio de transcripción y UTR (Ekblom & Wolf, 2014).

El aumento exponencial de datos de secuenciación disponibles que posibilitó la tecnología NGS requirió el desarrollo de métodos informáticos y estadísticos para procesar ese volumen de información y obtener datos útiles. Sin embargo, aún no ha sido posible automatizar la generación de datos de calidad. Por ejemplo, la velocidad de crecimiento de secuencias disponibles es mayor que la de curación de esa información que es depositada en bases de datos. Esto puede verse en forma gráfica al comparar el contenido de secuencias de la base de datos UniProt entre aquellas curadas manualmente (SwissProt) respecto de las incorporadas automáticamente (UniProt50) (The UniProt Consortium, 2017) (Fig. 1). Además, mucha de la información se publica sin descripción funcional y estructural para genomas, genes y proteínas. Los genomas completos son un claro ejemplo de esto, ya que, si bien aumentan día a día, al igual que sucede con el resto de los datos generados en grandes cantidades, la mayoría se encuentran poco curados, y muchos tienen poca o nula anotación o información verificada experimentalmente. Una de las razones de esto es que la velocidad con la que se producen los datos biológicos supera ampliamente la velocidad con la que se pueden llevar a cabo experimentos que den información biológica. La carencia de un conocimiento acabado de la estructura y contenido de los genomas representa un problema importante para la gran cantidad de análisis subsiguientes. De esta manera, el cuello de botella en los proyectos genómicos continúa siendo el análisis y procesamiento bioinformático de los datos provenientes de los secuenciadores, tanto en la etapa de ensamblado como en la anotación estructural y funcional.

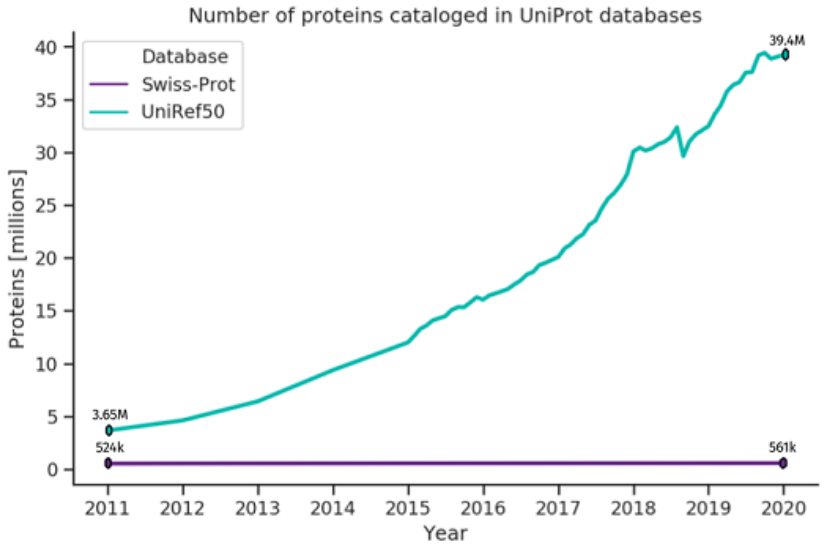

Fig. 1 Número de proteínas catalogadas en las bases de datos UniProt. SwissProt contiene proteínas revisadas y anotadas manualmente. Su crecimiento es imperceptible en comparación con UniRef50 que comprende secuencias no revisadas y anotadas automáticamente (The UniProt Consortium, 2017)

Como parte de un proyecto de postdoctorado en el Laboratorio Mixto de Biotecnología Acuática se obtuvieron los ensamblados correspondientes a genomas de referencia para una hembra y un macho de pacú a partir de datos de secuencias de tecnología Illumina. El ensamblado del genoma del macho se encuentra disponible bajo el número de acceso PRJNA754035 en NCBI (Mascali et al., 2022) (genoma de la hembra no publicado). Luego de la etapa de ensamblado, se inició la anotación *ab initio*, para lo cual se utilizó el programa AUGUSTUS (Stanke et al., 2006) para la predicción de regiones codificantes. Dado que ninguna estrategia de anotación en forma individual es capaz de anotar en forma completa el genoma, es recomendable utilizar distintas estrategias y herramientas para poder cubrir todos los casos. Entre ellas la anotación *ab initio*, la complementación con datos de transcriptómica y de otras especies, además de la tradicional comparación por homología con bases de datos. Estos enfoques son necesarios para el mejor aprovechamiento posible de los datos genómicos recientemente obtenidos para pacú.

OBJETIVOS

Objetivo general:

Contribuir en la anotación de los genomas ensamblados de macho y hembra de pacú *Piaractus mesopotamicus* utilizando los datos de transcriptómica disponibles en bases de datos.

Objetivos específicos:

Para cumplir con el objetivo general se proponen los siguientes objetivos específicos:

1. Evaluar la calidad de las de secuencias crudas de transcriptómica de pacú, disponibles en repositorios específicos.
2. Analizar la calidad de los ensamblados de transcriptoma de pacú realizados a partir de las secuencias crudas.
3. Describir la anotación de los transcriptomas ensamblados.
4. Describir la anotación de los genomas (macho y hembra) de pacú.

MATERIALES Y MÉTODOS

1. Obtención de secuencias disponibles de transcriptómica de pacú.

Los datos de transcriptoma de músculo (Mareco et al., 2015) fueron descargados como secuencias crudas (*reads*) del European Nucleotide Archive bajo el número de acceso PRJEB6656. Este proyecto consiste en 10 bibliotecas *paired-end* de 100 pb de longitud, 5 de músculo lento y 5 de músculo rápido, secuenciadas por Illumina HiSeq 2000. Por otro lado, los datos de transcriptoma de hígado fueron puestos a disposición por parte de los colaboradores del Centro de Aquicultura da UNESP, Universidade Estadual Paulista, Jaboticabal, SP, Brazil. Se encuentran disponibles otras secuencias de hígado bajo el número de acceso PRJNA632934 pero las mismas no fueron abordadas por una limitante de tiempo para realizar este trabajo.

Ambos conjuntos de datos fueron importados dentro del servidor FinisTerrae II del CESGA (Red Española de Supercomputación) para realizar allí el análisis de los datos.

2. Evaluación de la calidad de las lecturas

La calidad de las lecturas obtenidas de la base de datos del ENA fue evaluada utilizando los programas FastQC v0.11.7 (Andrews, 2010) y MultiQC v1.9 (Ewels et al., 2016) con parámetros estándar.

3. Ensamblado

Previo al ensamblado, los adaptadores y las secuencias crudas de baja calidad se eliminaron utilizando el programa Trimmomatic v0.38 (Bolger et al., 2014) con los siguientes parámetros: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36. Donde SLIDINGWINDOW realiza un recorte de ventana deslizante (de 4pb), cortando una vez que la calidad promedio dentro de la ventana cae por debajo de un umbral (20); MINLEN descarta la lectura si está por debajo de una longitud específica (36pb); LEADING y TRAILING cortan las bases del comienzo y al final de una lectura (respectivamente), si está por debajo de un umbral de calidad indicado (3).

El ensamblado de las lecturas de transcriptoma de músculo se realizó utilizando el programa Trinity (Haas et al., 2013), usando sólo las secuencias apareadas, con los siguientes parámetros: --seqTypefq --max_memory 60G --CPU 12--no_version_check --full_cleanup --min_contig_length 300. Las estadísticas del ensamblado se obtuvieron con la herramienta TrinityStats de Trinity.

La calidad del ensamblado se evaluó con las herramientas BUSCO 4.1.4 (Manni et al., 2021) y Bowtie2 2.3.5.1 (Langmead & Salzberg, 2012). BUSCO intenta proporcionar una evaluación cuantitativa de la integridad en términos del contenido

genético esperado de un ensamblaje de transcriptoma. Se utilizó la base de genes actinopterygii_odb10 (peces óseos) (fecha de creación: 2020-08-5, 3640 genes) y el modo “transcriptoma”. Con Bowtie2 se analizó la tasa de representación de las lecturas en el ensamblado final.

Finalmente, se limitó la redundancia en los transcriptos ensamblados usando CD-HIT-EST v4.8.1 (Fu et al., 2012) para colapsar *contigs* (región consenso de ADN) con al menos un 90 % de identidad (por ejemplo, aquellos alelos alternativos del mismo *locus*).

4. Anotación del transcriptoma obtenido.

La anotación funcional estuvo enfocada en anotar ARNm, debido a que en la construcción de las bibliotecas de secuenciación se utilizan métodos para aislar y enriquecer este tipo de ARN y evitar así la abundancia de ARNr. La anotación se realizó con los siguientes pasos:

* Identificación de los ORFs y predicción de la secuencia de aminoácidos.
* Transferencia de homología y asignación de identidad mediante búsqueda de secuencias.
* Anotación de características de secuencia.
* Ontología génica (GO)

4.1. Identificación de los ORFs y predicción de la secuencia de aminoácidos

Primero, se predijeron las regiones codificantes de los transcriptos ensamblados utilizando TransDecoder v. 5.5.0 (Haas et al., 2013) con la configuración predeterminada, seleccionando un único marco de lectura abierto (ORF) por transcripto de más de 100 aminoácidos. Este programa predice en base a modelos probabilísticos que tienen en cuenta la composición de nucleótidos y la longitud de los marcos de lectura abiertos, evaluando qué regiones podrían ser codificantes y, posteriormente, las traduce en secuencias de aminoácidos.

4.2. Transferencia de homología y asignación de identidad mediante búsqueda de secuencias.

La transferencia de homología puede considerarse la forma más básica de anotación transcriptómica. En este método, las secuencias ensambladas (como ADN/ARN o traducidas a proteínas) se suministran a las herramientas de búsqueda de secuencias como *queries* (consulta) y como *targets* (objetivo) se proporciona una base de datos de secuencias de referencia bien anotadas. A partir de las proteínas predichas por el TransDecoder se hizo una búsqueda por homología (blastp) usando Diamond (Buchfink et al., 2015) con la que se alineó cada una de las secuencias *query* con la base de datos no redundante de proteínas de SwissProt (Release 2022_02). La base de datos a emplear fue creada con la herramienta makedb, del programa Diamond. Se recuperó un solo resultado por secuencia *query*. También se realizó una

búsqueda de homología usando los transcriptos filtrados por CD_HIT_EST contra la base de datos de proteínas de SwissProt (blastx), recuperando un solo resultado por secuencia *query*.

4.3. Anotación de características de secuencia.

Se realizó la búsqueda de péptidos señal con el programa SignalP 6.0 (Dyrløv Bendtsen et al., 2004), de regiones transmembrana usando TMHMM 2.0 (Krogh et al., 2001) y el de motivos conservados con HMMER v3.3.2 (Finn et al., 2011), usando como entrada las predicciones del TransDecoder.

4.4. Ontología génica (GO)

Toda esta información se compiló con el programa Trinotate (Grabherr et al., 2011), que realiza la anotación funcional de transcriptomas *de novo* e integra esa información en una base de datos SQLite, que permite una rápida consulta a través de la interfaz gráfica de Trinotate Web.

5. Anotación del genoma

En el Laboratorio Mixto de Biotecnología Acuática se obtuvieron los ensamblados correspondientes a genomas de referencia para una hembra y un macho de pacú a partir de datos de secuencias de tecnología Illumina. El ensamblado del genoma del macho se encuentra disponible bajo el número de acceso PRJNA754035 en NCBI (Mascali, et al. 2022), mientras que el genoma de la hembra no publicado

Para anotar los genes que codifican proteínas en el genoma del pacú, explotamos enfoques *ab initio*, basados en RNA-seq y basados en homología, usando el programa MAKER v3.01.04 (Cantarel et al., 2008). El modelado de genes se realizó utilizando algoritmos de predicción de genes *ab initio* con evidencia de proteínas y transcriptomas mediante los procedimientos EST2GENOME y PROTEIN2GENOME en MAKER. El transcriptoma de músculo previamente ensamblado y agrupado en *clusters* se proporcionó como evidencia de EST, y como evidencia de homología de secuencias proteicas se seleccionaron de la base datos de secuencias de proteínas SwissProt aquellas provenientes de la clase *Actinopterygii*. Las predicciones *ab initio* de genes, realizadas con el programa AUGUSTUS también fueron incorporadas. Estos datos se usaron para anotar un genoma de macho y uno de hembra, en simultáneo.

El programa MAKER es controlado por tres archivos: maker_opts.ctl, maker_bopts.ctl y maker_exe.ctl. maker_exe.ctl contiene las rutas hacia los ejecutables que precisa el programa; maker_bopts.ctl permite modificar los parámetros del BLAST y, por último, maker_opts.ctl es donde radican la mayor parte de parámetros controladores del programa. En este trabajo, sólo se modificaron parámetros de este último archivo. Como ejemplo se muestra el archivo de corrida del anotado del genoma de macho (contigs_197):

```
#-----Genome (these are always required)
genome=MAKER_musculo/contigs_197.fa #genome sequence (fasta file or fasta embeded
in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic

#-----Re-annotation Using MAKER Derived GFF3
maker_gff= #MAKER derived GFF3 file
est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no
altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no
protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no
rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no
model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no
pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no
other_pass=0 #passthrough anything else in maker_gff: 1 = yes, 0 = no

#-----EST Evidence (for best results provide a file for at least one)
est=MAKER_musculo/musculo_cdhit.fasta #set of ESTs or assembled mRNA-seq in fasta
format
altest= #EST/cDNA sequence file in fasta format from an alternate organism
est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file
altest_gff= #aligned ESTs from a closely related species in GFF3 format

#-----Protein Homology Evidence (for best results provide a file for at least one)
protein=MAKER_musculo/sprot_actinoperygii.fa  #protein sequence file in fasta format
(i.e. from multiple organisms)
protein_gff=  #aligned protein homology evidence from an external GFF3 file

#-----Repeat Masking (leave values blank to skip repeat masking)
model_org= #select a model organism for RepBase masking in RepeatMasker
rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker
repeat_protein=/mnt/netapp1/Optcesga_FT2_RHEL7/2020/software/MPI/gcc/system/openmp
i/4.1.4_ft3/maker/3.01.04/data/te_proteins.fasta   #provide   a   fasta   file   of
transposable element proteins for RepeatRunner
rm_gff= #pre-identified repeat elements from an external GFF3 file
prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 =
yes, 0 = no
softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust
filtering)

#-----Gene Prediction
snaphmm= #SNAP HMM file
gmhmm= #GeneMark HMM file
augustus_species=zebrafish #Augustus gene prediction species model
fgenesh_par_file= #FGENESH parameter file
pred_gff=MAKER_musculo/augustusFiles/contigs_197_augustus.gff3          #ab-initio
predictions from an external GFF3 file
model_gff= #annotated gene models from an external GFF3 file (annotation pass-
through)
run_evm=0 #run EvidenceModeler, 1 = yes, 0 = no
est2genome=1 #infer gene predictions directly from ESTs, 1 = yes, 0 = no
protein2genome=1 #infer predictions from protein homology, 1 = yes, 0 = no
trna=0 #find tRNAs with tRNAscan, 1 = yes, 0 = no
snoscan_rrna= #rRNA file to have Snoscan find snoRNAs
snoscan_meth= #-O-methylation site fileto have Snoscan find snoRNAs
unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 = yes, 0 =
no
allow_overlap=0 #allowed gene overlap fraction (value from 0 to 1, blank for default)

#-----Other Annotation Feature Types (features MAKER doesn't recognize)
```

```
other_gff= #extra features to pass-through to final MAKER generated GFF3 file

#-----External Application Behavior Options
alt_peptide=C #amino acid used to replace non-standard amino acids in BLAST databases
cpus=1 #max number of cpus to use in BLAST and RepeatMasker (not for MPI, leave 1
when using MPI)

#-----MAKER Behavior Options
max_dna_len=300000 #length for dividing up contigs into chunks (increases/decreases
memory usage)
min_contig=1000 #skip genome contigs below this length (under 10kb are often useless)

pred_flank=200 #flank for extending evidence clusters sent to gene predictors
pred_stats=0 #report AED and QI statistics for all predictions as well as models
AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)
min_protein=0 #require at least this many amino acids in predicted proteins
alt_splice=0 #Take extra steps to try and find alternative splicing, 1 = yes, 0 =
no
always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 = no
map_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes, 0 =
no
keep_preds=0 #Concordance threshold to add unsupported gene prediction (bound by 0
and 1)

split_hit=10000 #length for the splitting of hits (expected max intron size for
evidence alignments)
min_intron=20 #minimum intron length (used for alignment polishing)
single_exon=0 #consider single exon EST evidence when generating annotations, 1 =
yes, 0 = no
single_length=250 #min length required for single exon ESTs if 'single_exon is
enabled'
correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes

tries=2 #number of times to try a contig if there is a failure for some reason
clean_try=0 #remove all data from previous run before retrying, 1 = yes, 0 = no
clean_up=0 #removes theVoid directory with individual analysis files, 1 = yes, 0 =
no
TMP= #specify a directory other than the system default temporary directory for
temporary files
```

Los mismos parámetros fueron usados también para el anotado del genoma de hembra (contigs_187).

Los datos de salida fueron integrados en un único archivo de formato gff para cada genoma, usando las herramientas fasta_merge y gff3_merge:

```
fasta_merge -d contigs_197_master_datastore_index.log
gff3_merge –n -d contigs_197_master_datastore_index.log
```

Para realizar el análisis de calidad del anotado se usó la herramienta AED_cdf_generator.pl para obtener las estadísticas del parámetro AED (Annotation Edit Distance) y se corrió el siguiente script para obtener el número de genes anotados y su largo promedio:

```
cat contigs_197.all.gff | awk '{ if ($3 == "gene") print $0 }' | awk '{ sum += ($5
- $4) } END { print NR, sum / NR }' > count_genes.txt
```

Como visualizador se utilizó el programa JBrowse 2 (Skinner, et al. 2009). JBrowse es un navegador de genomas que es capaz de visualizar diversos tipos de datos localizándolos en el genoma de referencia. Es capaz de integrar múltiples archivos. Se corrieron los siguientes comandos:

```
samtools faidx contigs_187.fa
sudo jbrowse add-assembly contigs_187.fa --out /var/www/html/jbrowse2 --load copy
gt gff3 -sortlines -tidy -retainids contigs_187.all.gff >
contigs_187.all.sorted.gff
bgzip contigs_187.all.sorted.gff
tabix contigs_187.all.sorted.gff.gz
sudo jbrowse add-track contigs_187.all.sorted.gff.gz --assemblyNames=contigs_187 --
-out /var/www/html/jbrowse2 --load copy
```

Luego, en un navegador se accedió al puerto local: http://localhost/jbrowse2

RESULTADOS Y DISCUSIÓN

1. Obtención de secuencias disponibles de transcriptómica de pacú.

Con el objetivo de anotar los genomas de pacú utilizando datos de transcriptomas, este trabajo inició con una búsqueda bibliográfica de los transcriptomas disponibles. Se pudieron identificar dos experimentos, uno con secuencias de ARNs provenientes de hígado (Mastrochirico-Filho et al., 2016) y otro de músculo (Mareco et al., 2015). El transcriptoma de hígado ensamblado fue cedido por el grupo de investigación del Dr. Hashimoto de la universidad UNESP de Brasil. Por otro lado, en cuanto al ensayo de músculo el transcriptoma ensamblado no estaba disponible, pero si fue posible descargar los datos crudos de las secuencias de ARN para las 10 muestras de la base de datos ENA (PRJEB6656). Estos datos requirieron trabajo posterior para su ensamblado.

2. Análisis de calidad del ensamblado de hígado

Como el ensamblado de este transcriptoma fue realizado por otro grupo, fue necesario evaluar la utilidad de los datos proporcionados. Para ello, se realizaron los mismos análisis de calidad que se usaron para el transcriptoma de músculo.

Su información básica indica un total de 4110 transcriptos y un %GC = 46,65. Un parámetro cuantitativo útil para describir los resultados es el valor de N50. Este es una medida de la calidad de ensamblaje de los datos NGS. Se define como una estadística mediana ponderada tal que el 50% de todo el conjunto está contenido en *contigs* que son iguales o mayores que este valor Las estadísticas basadas en todos los transcriptos arrojaron un N50 de 871 pb y un largo promedio de los *contigs* de 804,74 pb. La búsqueda de ortólogos con BUSCO, comparando contra el conjunto de *Actinopterygii* presentó los siguientes resultados:

Tabla 1 Estadísticas BUSCO

	N° genes	Porcentaje
Complete BUSCOs (C)	151	4,2%
Complete and single-copy BUSCOs (S)	148	4,1%
Complete and duplicated BUSCOs (D)	3	0,1%
Fragmented BUSCOs (F)	73	2%
Missing BUSCOs (M)	3416	93,8%

Si un ensamblado tiene una alta proporción de genes BUSCO faltantes o fragmentados, es de mala calidad. Por esto, el transcriptoma de hígado no fue tenido en cuenta en análisis posteriores.

3. Ensamblado del Transcriptoma de músculo.

3.1. Evaluación de la calidad de las secuencias crudas de musculo.

Para el ensamblado del transcriptoma de músculo, la calidad de las secuencias crudas de músculo se evaluó con la herramienta FastQC y se reunieron los resultados para facilitar su análisis mediante el programa MultiQC. En promedio, cada muestra tenía 76,91 millones de lecturas, de las cuales más del 80% estaban duplicadas (Tabla 2). La calidad de las secuencias fue mayor a 30 en casi toda su extensión (Fig. 2), indicando una alta calidad de secuenciación. Estos parámetros indican una buena calidad de los datos crudos, por lo que se continuó el análisis usando todas las muestras.

Tabla 2: Resumen de las características de las muestras de transcriptoma de musculo.

SampleName	% Dups	% GC	M Seqs
ERR556963	82.8%	48%	73.1
ERR556964	81.0%	48%	85.7
ERR556965	82.9%	48%	83.2
ERR556966	84.1%	47%	82.4
ERR556967	84.9%	48%	68.0
ERR556968	88.2%	50%	90.5
ERR556969	87.1%	50%	69.8
ERR556970	86.1%	51%	72.4
ERR556971	87.5%	50%	72.5
ERR556972	87.5%	50%	71.5

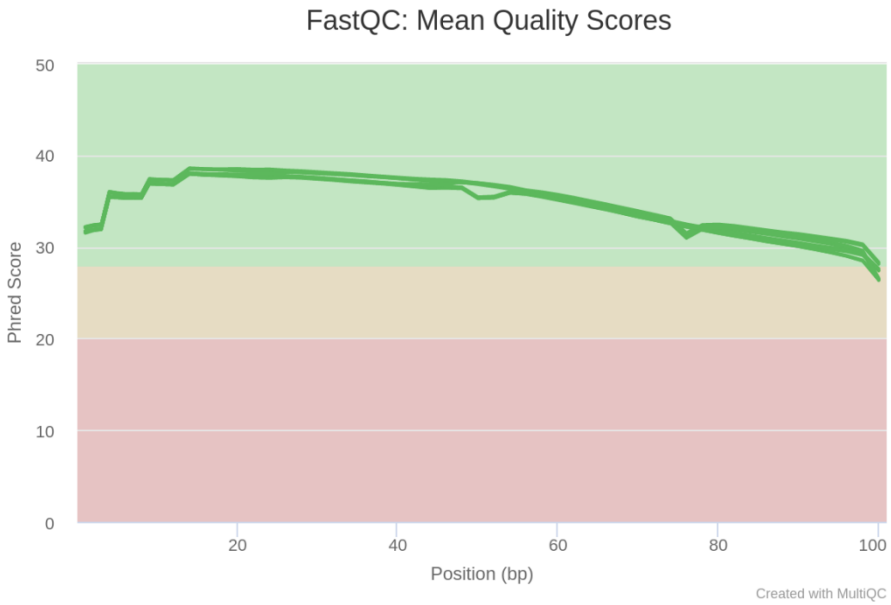

Fig. 2 Superposición de gráficos de calidad de secuencia para las 10 muestras del experimento de transcriptoma de músculo. En el eje horizontal se indica la posición en pb dentro de cada lectura, y en el eje vertical el valor del parámetro de calidad correspondiente.

Previo al ensamblado, se quitaron los adaptadores y secuencias de baja calidad utilizando la herramienta Trimmomatic, recuperando el 92,8% de las secuencias. Posteriormente, las 5 bibliotecas de músculo rápido y las 5 de músculo lento se ensamblaron con el programa Trinity de manera conjunta para generar un único transcriptoma consenso. El programa devuelve como resultado los transcriptos ensamblados y también agrupados en “genes”. Un gen Trinity es una colección de transcriptos relacionados. En un ensamblado *de novo*, los transcriptos se relacionan por haber compartido *kmers* (subcadena de longitud k) durante la etapa de ensamblado. Como resultado se obtuvieron un total de 144.712 “genes” Trinity, 231.914 transcriptos y un %GC = 44,44. Para describir los resultados puede usarse el valor de N50. Sin embargo, en transcriptomas, a menudo estos valores se pueden exagerar debido a que el programa genera demasiadas isoformas, especialmente para los transcriptos más largos. Para mitigar este efecto, se calculan los valores de N50 basándose en el uso sólo de la isoforma más larga por 'gen'. Las estadísticas usando esa estrategia arrojaron un N50 de 1236 pb y un largo promedio de los contigs de 884,83 pb. Estos resultados son similares a los publicados por los autores del trabajo original y con otros trabajos de ensamblado de transcriptomas de peces (Carruthers et al., 2018). A continuación, se muestra un resumen de los resultados (Tabla 3):

Tabla 3 Resumen de estadísticas finales del ensamblado.

	Estadísticas basadas en TODOS los *contigs*	**Estadísticas basadas SOLO en la ISOFORMA MAS LARGA de cada 'GEN'**
Contig N10	6195	5073
Contig N20	4580	3528
Contig N30	3606	2539
Contig N40	2877	1787
Contig N50	2266	1236
Mediana del largo de contig	665	507
Largo promedio	1273,31	884,83
Total de bases ensambladas	295.297.720	128.045.499

3.2. Control de calidad del ensamblado de músculo.

La calidad de un ensamblado se puede evaluar desde varias perspectivas. Primero está la longitud de la secuencia y la fragmentación. Un ensamblaje con muchos *contigs* cortos puede considerarse fragmentado. Es posible que esto sea el resultado de un ensamblaje inadecuado o una secuenciación deficiente. Se pueden usar herramientas para calcular estadísticas de longitud de secuencia (como el valor N50), evidenciadas en la sección anterior. En segundo lugar, se encuentra el soporte de las *reads*, es decir, la fracción de todas las lecturas que fueron utilizadas para el

ensamblado. Un ensamblado de buena calidad habrá hecho uso de la mayoría de las lecturas. Esta métrica se puede verificar fácilmente alineando las lecturas con las secuencias ensambladas, mientras mayor cantidad de lecturas alineen contra el transcriptoma mayor será la calidad del ensamblado. Para ello se utilizó la herramienta Bowtie, y el porcentaje de lecturas alineadas sobre el ensamblado fue de 98,84 %, indicando un alto aprovechamiento de todas las lecturas.

Otro método que sirve para evaluar tanto la calidad como la integridad es analizar la presencia de ortólogos de genes que son universales, se expresan persistentemente y se presentan casi exclusivamente como copias únicas en el genoma. Si un transcriptoma se ha secuenciado y ensamblado correctamente, se deberían encontrar ortólogos para la gran mayoría de estos genes. Este análisis se realizó con la herramienta BUSCO (Benchmarking Universal Single-Copy Orthologs) comparando al transcriptoma ensamblado con un conjunto de 3640 genes del grupo de organismos *Actinopterygii*. Los resultados se dividen en las categorías CS (completos y de copia simple), CD (completos y duplicados) F (fragmentados) M (ausentes), y se detallan a continuación:

Tabla 4 Estadísticas BUSCO

	N° genes	Porcentaje
Complete BUSCOs (C)	2961	81,4%
Complete and single-copy BUSCOs (CS)	1404	38,6%
Complete and duplicated BUSCOs (CD)	1557	42,8%
Fragmented BUSCOs (F)	149	4,1%
Missing BUSCOs (M)	530	14,5%

Como regla general, en un ensamblado de buena calidad al menos el 80 % de los genes analizados por BUSCO deben tener coincidencias en el transcriptoma. En general, los ensamblados de transcriptomas *de novo* tendrán muchas coincidencias duplicadas con las secuencias de BUSCO. Esto se debe a la presencia de transcriptos estrechamente relacionados que representan isoformas de corte y empalme y, por lo tanto, no es necesariamente indicativo de una redundancia no deseada en el ensamblado (Raghavan et al., 2022). Tomando esta información en consideración, el ensamblado presenta altos niveles de calidad.

3.3. Agrupación en clusters y reducción de la redundancia.

Los ensambladores de transcriptoma *de novo* suelen producir muchas más secuencias de las que cabría esperar en función del número de genes en el genoma, debido principalmente a las distintas isoformas de cada gen. Esto no es la excepción en nuestro caso, como podemos interpretar del valor de 42,8% de genes duplicados en los parámetros de calidad de BUSCO. Para simplificar este ensamblado, se utilizó la herramienta de agrupación en clústeres CD-HIT, con un umbral de identidad de secuencia del 90%, agrupando secuencias en grupos y extrayendo las secuencias

representativas. Los representantes son la secuencia más larga de cada grupo o la secuencia con mayor similitud con los miembros del grupo. Este punto de corte ha sido usado por otros autores previamente (Chabikwa et al., 2020).

Como resultado se obtuvieron un total de 144.582 “genes” Trinity, 181.619 transcriptos y un %GC = 43,75. Esto representa una reducción al 78,3% de los transcriptos, pero solo al 99,9% de los que Trinity considera genes. Las estadísticas basadas únicamente en la isoforma más larga de cada “gen” arrojaron un N50 de 1237 pb y un largo promedio de los contigs de 885,29 pb, prácticamente iguales a lo anterior. Este agrupamiento eliminó isoformas pero mantuvo la integridad general del transcriptoma. Es sabido que la cantidad de genes en organismos similares ronda los 30.000-40.000 genes (Carruthers et al., 2018), por lo que es posible que persistan isoformas y ARN no codificantes.

4. **Anotación funcional del transcriptoma de músculo.**

Luego de ensamblar y controlar la calidad del transcriptoma, procedimos a estudiar sus secuencias para anotarlas, es decir, dilucidar la funcionalidad que representan, y así poder utilizarlas posteriormente en el anotado del genoma. Para ello lo primero que hicimos fue una predicción *ab initio* de ORFs utilizando un programa que se basa en modelos probabilísticos. Estos ORFs luego serán útiles como conjuntos de prueba para para notar los CDS. Usando como *input* las secuencias agrupadas en clústeres, TransDecoder predijo 57.772 secuencias peptídicas.

A continuación, se realizó una búsqueda de secuencias de bases de datos homólogas a las nuestras para realizar la transferencia de homología. Se decidió usar la base de datos SwissProt, porque al estar curada manualmente se la considera la base de datos de mayor calidad. Como *queries* se usaron tanto los 181.619 transcriptos (blastx: secuencias traducidas contra una base de datos de secuencias de proteínas) como las 57.772 proteínas obtenidas por el predictor *ab initio* TransDecoder (blastp: secuencias de proteínas contra secuencias de proteínas). Las búsquedas se realizaron con el programa Diamond, que está orientado exclusivamente a la búsqueda en bases de datos de proteínas. Diamond es tan sensible como blastp y es 80 veces más rápido (Raghavan et al., 2022). En cada búsqueda se recuperó el mejor alineamiento. Para las secuencias obtenidas por TransDecoder, se obtuvieron 43.244 resultados y al realizar la búsqueda con los transcriptos, se obtuvieron 50.178 coincidencias.

Posteriormente, se anotaron dominios específicos de las proteínas utilizando programas que tienen en cuenta no solo la estructura primaria sino también su estructura secundaria. De manera particular, se buscaron péptidos señal utilizando SignalP, encontrándose 3.774 coincidencias. Por otro lado, se buscaron dominios transmembrana con TMHMM y se predijo al menos una hélice transmembrana en el 16,95% de las secuencias. De manera general, se utilizó el programa hmmscan

(HMMER) para contrastar contra la base de datos Pfam para la identificación de dominios proteicos y familias. Este otorgó 541.130 resultados.

Todos estos resultados fueron reunidos por el programa Trinotate. Éste, a su vez, suma las anotaciones GO a través de asignaciones transitivas de los principales resultados de búsqueda de homología. Los datos de anotaciones funcionales derivados del análisis de transcripciones se integraron en una base de datos SQLite que permite una búsqueda rápida y eficiente. Los resultados son entonces visibles a partir de una interfaz gráfica de usuario (GUI) TrinotateWeb (Fig. 3 y Fig. 4). Finalmente, el 35,3% (64.111 secuencias) de los transcriptos ensamblados pudo obtener algún tipo de anotación con este método. Este análisis indica que el uso de las secuencias del transcriptoma de músculo puede ser muy informativo en el proceso de anotado de los genomas

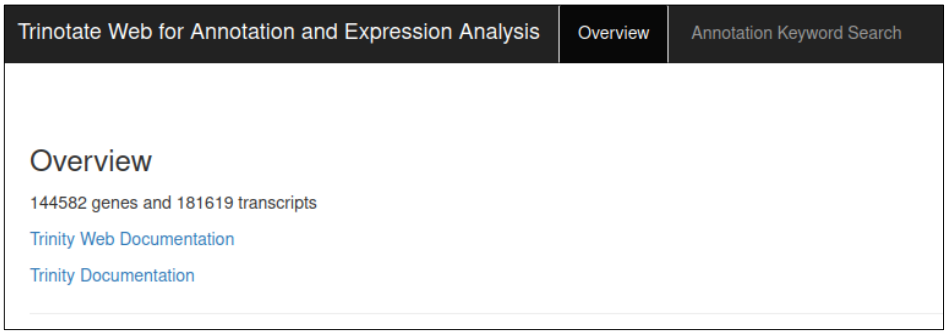

Fig. 3 Página de portada de Trinotate Web. Allí se indican la cantidad de transcriptos y las anotaciones realizadas.

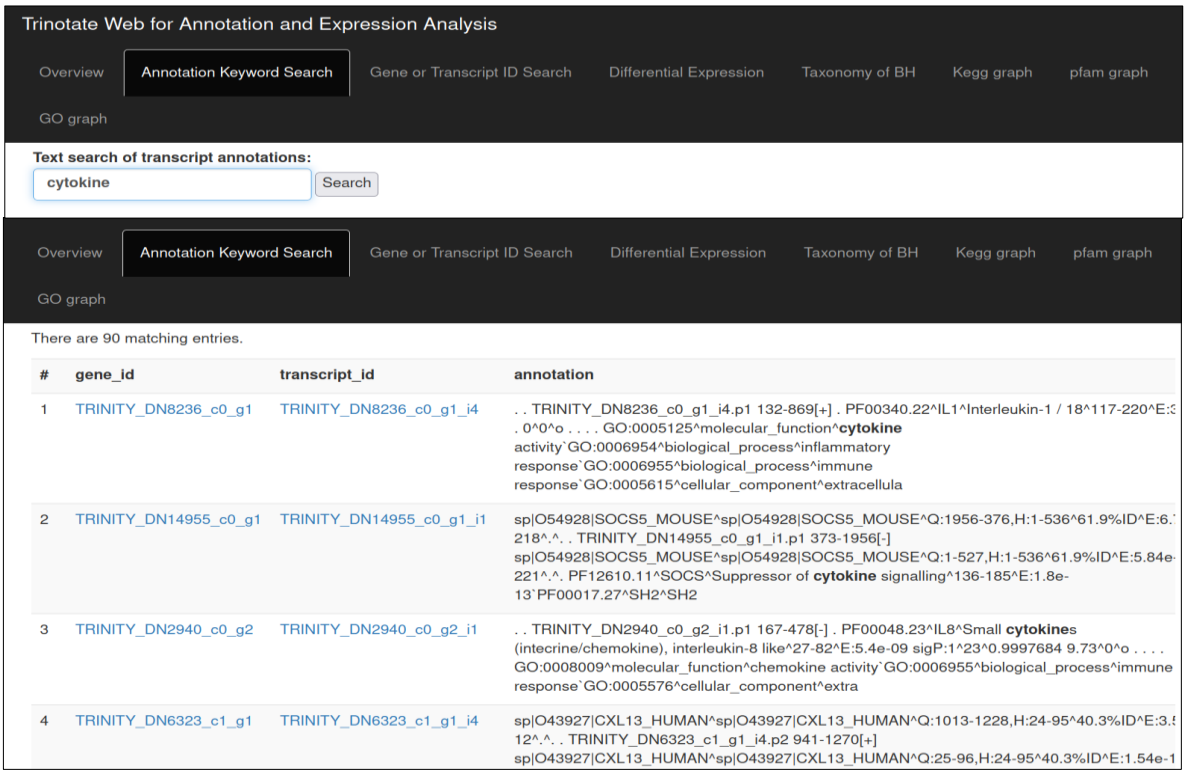

Fig. 4 Trinotate Web. Ejemplo de búsqueda del término de anotación “cytokine” y los resultados generados.

5. Anotación de los genomas de pacú

El objetivo último de este trabajo es contribuir a la anotación de dos genomas de pacú recientemente ensamblados por nuestro grupo de investigación. De manera resumida se muestran las principales estadísticas de los mismos en la Tabla 5:

Tabla 5 Principales estadísticas de los ensamblados de genoma de pacú macho y hembra.

Estadísticas	Macho (contigs_197)	Hembra (contigs_187)
# contigs	372.987	269.281
# contigs (>= 1000 bp)	168.755 (45%)	110.109 (41%)
Largest contig (bp)	219.307	328.659
Total length (bp)	1.252.140.284	1.246.951.483
N50	11.470	21.548
GC (%)	39,45	39,37

Se trabajó solo sobre aquellos *contigs* mayores a 1000 pb (aproximadamente el 45% de los datos de cada ensamblado). Se decidió este punto de corte, aunque la bibliografía sugiera descartar secuencias de menos de 10 Kb por el alto nivel de fragmentación de los genomas (Cantarel et al., 2008). De usar el punto de corte recomendado se perdería la mayor parte de los genomas y por ende su información.

Los análisis fueron realizados usando el programa MAKER v3.01.04. Para anotar los genomas se utilizaron los transcriptos ensamblados y agrupados en *clusters*, las predicciones *ab initio* de genes realizadas con el programa AUGUSTUS y secuencias de proteínas de la base de datos SwissProt.

En cuanto a estas últimas, se hizo una selección de secuencias de proteínas provenientes de la clase *Actinopterygii* dentro la base de datos de SwissProt (5.542 secuencias). Si bien son pocas secuencias, se hizo esta selección porque SwissProt está manualmente curado por expertos. Contiene descripciones de proteínas, incluida la función, la estructura del dominio, la ubicación subcelular, las modificaciones postraduccionales y las variantes funcionalmente caracterizadas. Es un enfoque conservador, que espera obtener pocas anotaciones de alta calidad. Se decidió este enfoque porque MAKER permite construir sucesivas anotaciones usando la información previamente obtenida. De esta manera, se puede realizar este enfoque más estricto como primera aproximación e ir agregando bases de datos más grandes posteriormente.

El análisis fue realizado en el servidor FinisTerrae II del CESGA, el cual posee una cuota de uso de 3 Tb y 240.000 archivos. MAKER crea una jerarquía de capas de subdirectorios anidados y coloca los resultados de cada *contig* dentro de este sistema de archivos de miles de directorios anidados, lo cual permite un mejor rendimiento en el acceso a los resultados. Sin embargo, esto limitó el procesamiento de los genomas,

ya que al estar altamente fragmentados produjeron que se llegue al límite de cuota de archivos del servidor de manera diaria. Una forma de sobrepasar este inconveniente era ir comprimiendo en un único archivo un número de las carpetas generadas para liberar cuota de archivos. Sin embargo, el trabajo siguiendo esta metodología resultaba incómodo, susceptible a errores y la proyección de tiempo esperado para completar el análisis era de más de un mes para cada genoma. Por esta razón, desafortunadamente el procesamiento de los genomas no se pudo realizar en su totalidad.

El genoma de la hembra se alcanzó a procesar en un 38% (102.327 *contigs*), anotando 5.970 genes, con un largo promedio de 4.660,44 bases. El genoma del macho, en cambio, se procesó en un 23% (85.787 *contigs*), anotando 4.182 genes, con un largo promedio de 3.953,42 bases. Con estos resultados se puede notar el impacto tanto de las predicciones *ab initio* de AUGUSTUS como el aporte de las secuencias de transcriptoma para localizar genes, ya que se obtuvieron más genes que los que se esperaría obtener solo por homología de secuencia contra la base de datos SwissProt.

Para evaluar la calidad se usó el parámetro AED, que mide la bondad de ajuste de una anotación a la evidencia (predicciones *ab initio*, ESTs, etc) que la respalda. AED es un número entre 0 y 1, con un AED de cero que indica una concordancia perfecta con la evidencia disponible y un valor de uno que indica una ausencia total de soporte para el modelo de gen anotado (Eilbeck et al., 2009). En otras palabras, la puntuación AED proporciona una medida de la congruencia de cada gen anotado con su evidencia de respaldo. Idealmente, el 95% o más de los modelos de genes tendrán una AED de 0,5 o mejor en el caso de buenos ensamblados (Campbell et al., 2014). Para el caso del genoma de hembra, el 96,2% de los genes anotados tiene valores AED menores o iguales a 0,5. Para el genoma del macho, el porcentaje es del 96,4%. Los resultados se resumen en la Tabla 6.

Tabla 6 Resumen de resultados obtenidos con el programa MAKER.

	Hembra	Macho
Nivel de procesamiento (%)	38%	23%
Genes	5.970	4.182
Largo promedio (Campbell, et al.)	4.660,44	3.953,42
AED ≤ 0,5	96,2%	96,4%

Para visualizar los resultados se usó el programa JBrowse2. Como ejemplo se muestra un *contig* del genoma de hembra con el archivo de salida MAKER (formato gff) (Fig. 5). Así se observa como el programa integra la evidencia de los datos transcriptómicos con los obtenidos en la predicción *ab initio* de AUGUSTUS y los resultados de la búsqueda por homología contra la base de datos de SwissProt.

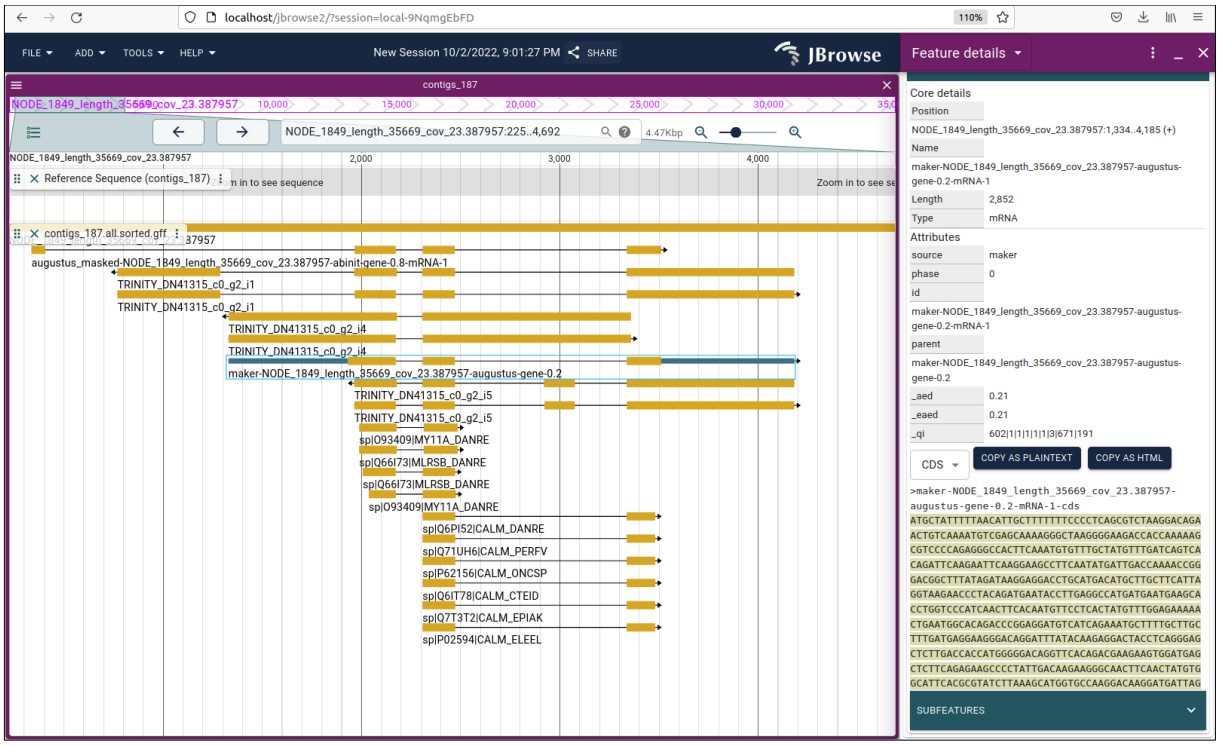

Fig. 5 JBrowse. Se observa la navegación por un contig y sus respectivas anotaciones. En amarillo el alineamiento con transcriptos, predicciones génicas y alineamientos contra la base de datos SwissProt. En verde y amarillo, un anotado consenso que realiza MAKER. Recuadro derecho: la descripción de las características de la región anotada por MAKER. Se observa “largo”, “AED”, CDS, entre otras características.

CONCLUSIÓN

Si bien contar con genomas de referencia para pacú (*P. mesopotamicus*) fue un gran avance para aumentar el conocimiento genómico sobre la especie, la falta de anotación de los mismos limita su aplicación. Es por ello que el presente trabajo intentó contribuir a la descripción de los genomas utilizando como recurso datos transcriptómicos de secuencias de pacú. Se comenzó analizando los datos disponibles para trabajar. Por un lado, se analizaron secuencias ensambladas y cedidas por un grupo colaborador correspondientes a muestras de hígado, pero las mismas fueron descartadas para el análisis posterior por la baja calidad que presentaba. Por otro lado, fueron descargadas de una base de datos pública secuencias crudas de un transcriptoma de músculo. Estas secuencias fueron filtradas y ensambladas generando un ensamblado de alta calidad, con un alto aprovechamiento de las secuencias crudas (98,84 %) y alta presencia (81,4%) de ortólogos de ciertos genes que son universales, se expresan persistentemente y se presentan casi exclusivamente como copias únicas en el genoma. A su vez, al anotar el transcriptoma se encontraron 64.111 genes e isoformas, indicando una perspectiva de alta utilidad en el uso de estas secuencias para anotar los genomas.

Los genomas obtenidos para un macho y una hembra de pacú a partir de tecnología Illumina se encuentran altamente fragmentados. Esto trajo como consecuencia una gran dificultad computacional en el proceso de anotación de los mismos, a pesar de trabajar en un servidor con un alto poder de cómputo. Por esto, solo fue posible la anotación de un fragmento de cada genoma. A pesar de ser resultados preliminares los mismos son alentadores, ya que presentan altos niveles de calidad según el parámetro AED. Algunas mejoras que se pueden realizar sobre el mismo son: completar el enmascaramiento de las regiones repetitivas (predichas con el programa RepeatModeler) usando el programa RepeatMasker; aumentar el punto de corte de las secuencias a analizar a 10 Kb, como sugiere el manual de MAKER, con el fin de reducir la complejidad computacional; y aunque sea computacionalmente más demandante, agregar una ronda de MAKER con una búsqueda de homología de secuencia contra una base de datos de proteínas más grande, como por ejemplo la fracción *Actinopterygii* de UniProt/TrEMBL (que contiene 6.107.807 secuencias).

BIBLIOGRAFÍA

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Babraham Bioinformatics.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, *48*(1). https://doi.org/10.1002/0471250953.bi0411s48

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Carruthers, M., Yurchenko, A. A., Augley, J. J., Adams, C. E., Herzyk, P., & Elmer, K. R. (2018). De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. *BMC Genomics*, *19*(1), 32. https://doi.org/10.1186/s12864-017-4379-x

Chabikwa, T. G., Barbier, F. F., Tanurdzic, M., & Beveridge, C. A. (2020). De novo transcriptome assembly and annotation for gene discovery in avocado, macadamia and mango. *Scientific Data*, *7*(1), 9. https://doi.org/10.1038/s41597-019-0350-9

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M., & Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research*, *7*. https://doi.org/10.12688/f1000research.13598.1

Dunham, R. A., Taylor, J. F., Rise, M. L., & Liu, Z. (2014). Development of strategies for integrated breeding, genetics and applied genomics for genetic improvement of aquatic organisms. *Aquaculture*, *420–421*, S121–S123. https://doi.org/10.1016/j.aquaculture.2013.10.020

Dyrløv Bendtsen, J., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology*, *340*(4), 783–795. https://doi.org/10.1016/j.jmb.2004.05.028

Eilbeck, K., Moore, B., Holt, C., & Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, *10*(1), 67. https://doi.org/10.1186/1471-2105-10-67

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. https://doi.org/10.1111/eva.12178

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl), W29–W37. https://doi.org/10.1093/nar/gkr367

Flores Nava, A. (2007). Aquaculture seed resources in Latin America: a regional synthesis. In M. Bondad-Reantaso (Ed.), *Assessment of Freshwater Fish Seed Resources for Sustainable Aquaculture* (pp. 91–102). FAO.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Honglang, H. (2007). 7.5 Freshwater fish seed resources. In M. Bondad-Reantaso (Ed.), *Assessment of Freshwater Fish Seed Resources for Sustainable Aquaculture* (pp. 185–199). FAO.

Huete-Pérez, J. A., & Quezada, F. (2013). Genomic Approaches in Marine Biodiversity and Aquaculture. *Biological Research*, *46*(4), 353–361. https://doi.org/10.4067/S0716-97602013000400007

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Laurenti, A. L. G., Bianchi, M. C. G., Chopin, F., Farme, T., Franz, N., Fuentevilla, C., & Garibaldi, L. (2014). *FAO: the state of world fisheries and aquaculture*.

Liotta, J. (2005). *Distribución geográfica de los peces de aguas continentales de la República Argentina*. ProBiota. FCNyM, UNLP.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Mareco, E. A., Garcia de la Serrana, D., Johnston, I. A., & Dal-Pai-Silva, M. (2015). Characterization of the transcriptome of fast and slow muscle myotomal fibres in the pacu (Piaractus mesopotamicus). *BMC Genomics*, *16*(1), 182. https://doi.org/10.1186/s12864-015-1423-6

Mascali, F. C., Posner, V. M., Romero Marano, E. A., del Pazo, F., Hermida, M., Sánchez, S., Mazzoni, T. S., Martinez, P., Rubiolo, J. A., & Villanova, G. V. (2022). Development and validation of sex-specific markers in Piaractus mesopotamicus. *Aquaculture*, *558*, 738374. https://doi.org/10.1016/j.aquaculture.2022.738374

Mastrochirico-Filho, V. A., Borges, C. H. S., Freitas, M. V., Ariede, R. B., Pilarski, F., Utsunomia, R., Carvalheiro, R., Gutierrez, A. P., Peñaloza, C., Yáñez, J. M., Houston, R. D., & Hashimoto, D. T. (2020). Development of a SNP linkage map and genome-

wide association study for resistance to Aeromonas hydrophila in pacu (Piaractus mesopotamicus). *BMC Genomics*, *21*(1), 672. https://doi.org/10.1186/s12864-020-07090-z

Mastrochirico-Filho, V. A., Hata, M. E., Kuradomi, R. Y., de Freitas, M. V., Ariede, R. B., Pinheiro, D. G., Robledo, D., Houston, R., & Hashimoto, D. T. (2020). Transcriptome Profiling of Pacu (Piaractus mesopotamicus) Challenged With Pathogenic Aeromonas hydrophila: Inference on Immune Gene Response. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00604

Mastrochirico-Filho, V. A., Hata, M. E., Sato, L. S., Jorge, P. H., Foresti, F., Rodriguez, M. V., Martínez, P., Porto-Foresti, F., & Hashimoto, D. T. (2016). SNP discovery from liver transcriptome in the fish Piaractus mesopotamicus. *Conservation Genetics Resources*, *8*(2), 109–114. https://doi.org/10.1007/s12686-016-0521-3

Pimentel, J. da S. M., Carmo, A. O. do, Maciel, D. de C. L., Siqueira, F. de F., & Kalapothakis, E. (2014). Complete mitochondrial genome sequence of *Piaractus mesopotamicus* (Holmberg, 1887). *Mitochondrial DNA*, 1–2. https://doi.org/10.3109/19401736.2014.971297

Purcell, C. M., Seetharam, A. S., Snodgrass, O., Ortega-García, S., Hyde, J. R., & Severin, A. J. (2018). Insights into teleost sex determination from the Seriola dorsalis genome assembly. *BMC Genomics*, *19*(1), 31. https://doi.org/10.1186/s12864-017-4403-1

Raghavan, V., Kraft, L., Mesny, F., & Rigerte, L. (2022). A simple guide to de novo transcriptome assembly and annotation. *Briefings in Bioinformatics*, *23*(2). https://doi.org/10.1093/bib/bbab563

Resende, E. K. (2003). Migratory fishes of the Paraguay-Paraná basin, excluding the Upper Paraná basin. In J. Carolsfield, B. Harvey, C. Ross, & A. Baer (Eds.), *Migratory Fishes of South America: biology, fisheries and conservation status.* (pp. 99–156). World Fisheries Trust, World Bank, IDRC.

Stanke, M., Tzvetkova, A., & Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology*, *7*(Suppl 1), S11. https://doi.org/10.1186/gb-2006-7-s1-s11

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174