INIDEP

INSTITUTO NACIONAL DE INVESTIGACIÓN Y DESARROLLO PESQUERO

INFORME DE INVESTIGACIÓN

Número	Páginas	Fecha de aprobación
062	12	31 de Agosto de 2020

Dirección
DIRECCIÓN DE PESQUERIAS DE INVERTEBRADOS Y AMBIENTE MARINO

Programa / Gabinete
Gabinete de Biología Molecular y Microbiología

Actividad
Estudio de la diferenciación sexual y la reproducción en peces de interés en acuicultura

SECUENCIACIÓN, ENSAMBLAJE *DE NOVO* Y ANOTACIÓN DEL GENOMA DEL LENGUADO NEGRO

El lenguado negro (*Paralichthys orbignyanus*) es un pez económicamente importante por su gran potencial de cultivo en Argentina. En este estudio se ha secuenciado, ensamblado y anotado el genoma completo de esta especie. La secuenciación fue realizada con un equipo HiSeq-4000 de Illumina. Se construyeron cuatro bibliotecas independientes (macho, hembra, pool de machos y pool de hembras); con lecturas pareadas de 150 pb (pares de bases), longitud media de inserción de 350 pb y coberturas mayores a 35 veces cada una. Se realizó un ensamblaje *de novo* de las secuencias obtenidas. Cada biblioteca se ensambló por separado utilizando SOAPdenovo2 y los *contigs* resultantes se ensamblaron con SSPACE3, antes de rellenar los gaps (huecos) con GapCloser. El tamaño del genoma obtenido fue de ~550 Mbp, tanto para machos como para hembras (41,35% del contenido de GC, 0,11% de bases indeterminadas). Un análisis de ensambles a nivel de genes centrales (subconjunto de 458 familias KOG expresadas universalmente) mostró más del 98% de los genes centrales presentes, con más del 78% de ellos, con una cobertura superior al 50%. Se realizó una predicción de genes basada en predictores estadísticos y predictores basados en secuencias de otra especie estrechamente relacionada, el *Paralichthys olivaceus*. Se encontraron 25,231 genes codificantes de proteínas, 445 ARNt, 3 ARNr y más de 1,500 ARNs no codificantes. Un mapa del genoma de esta especie ayudará en futuros estudios de genes relacionados con la diferenciación sexual, la pubertad, la ingesta de alimentos, la reproducción y otras funciones importantes relacionadas con el cultivo del lenguado negro.

Citar Indicando la fuente. El contenido no debe ser reproducido total o parcialmente sin la expresa conformidad del INIDEP

	Institución	Cargo
SOLICITADO POR	Marcelo Pájaro	Director de área

PREPARADO POR	**APROBADO POR**
Firma: Nombre: MECHALY, ALEJANDRO SEBASTIAN	María Inés Trucco Jefe de Programa / Gabinete
Firma: Nombre: y colaboradores	Director de área
	Dra. CLAUDIA RAQUEL CAROZZA DIRECCIÓN NACIONAL INVESTIGACIÓN INIDEP Director Nacional de Investigación
	Director del INIDEP

COPIA ELECTRONICA INIDEP

SECUENCIACIÓN, ENSAMBLAJE *DE NOVO* Y ANOTACIÓN DEL GENOMA DEL LENGUADO NEGRO

Alejandro S. Mechaly[1], German Burguener[2], Ezequiel Sosa[2], Fernando Villareal[3], Adrián Turjanski[2], Jordi Viñas[4] , Paulino Martinez[5]

[1]CONICET-INIDEP, Mar del Plata, Argentina;[2]UBA, Buenos Aires, Argentina;[3]IIB-CONICET, Mar del Plata, Argentina;[4]Universitat de Girona, Girona, España; [5]Universidad Santiago de Compostela, Lugo, España.

COPIA ELECTRONICA INIDEP

INTRODUCCION

El lenguado negro (*Paralichthy sorbignyanus*) es un pez marino que pertenece al orden Pleuronectiformes y su distribución está restringida a Sudamérica (Díaz de Astarloa, 2002). Este lenguado es un recurso pesquero importante en Brasil, Uruguay y Argentina y habita en zonas de estuarios poco profundos y aguas costeras desde Río de Janeiro, Brasil (Haimovici*et al*., 1996) hasta el Golfo de San Matías, Argentina (Díaz de Astarloa y Munroe, 1998; Díaz de Astarloa, 2002). En el INIDEP, se ha logrado cerrar de forma completa su ciclo reproductivo, obteniéndose la puesta a punto de las condiciones de cultivo (Bambill*et al*., 2006; López *et al*., 2009a, 2009b; Radonic*et al*., 2007; Radonic y Macchi, 2009; Boccanfuso *et al*., 2019). Si bien, en Argentina el lenguado negro tiene un gran potencial económico en acuicultura, todavía existen dificultades técnicas para obtener una producción a escala industrial. Dentro de los principales objetivos en toda industria encontramos la disponibilidad de tener un método de obtención de planteles todo hembra, que por su mayor tasa de crecimiento, son de crucial importancia en los cultivos a escala industrial. Por otro lado, en cultivo es importante disponer de conocimientos que permitan controlar el inicio de la pubertad y con ello evitar el gasto de energía en la reproducción. En este sentido, el conocimiento de métodos de producción de hembras y tener conocimientos que ayuden a controlar la reproducción en peces es crucial. Ambos aspectos, se han logrado controlar en otras especies de peces de cultivo acuícola, estudiando mayormente los genes que están involucrados en estos procesos.

Como ya mencionamos, se han realizaron varios estudios zootécnicos en esta especie durante las últimas dos décadas, sin embargo, existen pocos estudios genéticos. Por ejemplo, al buscar en la base de datos de GenBank, solo hay 12 secuencias de proteínas disponibles a Mayo del 2020. Aspectos importantes, como la determinación y diferenciación sexual, la proporción de sexos, la nutrición, la reproducción y otras, pueden estudiarse desde un punto de vista genético. Por lo tanto, disponer del genoma de esta especie es de gran ventaja para lograr estudios genéticos en esta especie. Durante la

última década, la revolución que ha experimentado la genómica en el mundo, ha ayudado a que actualmente existan una gran cantidad genomas secuenciados en peces. En lo que respecta a peces planos, actualmente existen disponible los genomas de cuatro especies; tonguesole (*Cynoglossussemilaevis*) (Chen*et al*., 2014), rodaballo (*Scophthalmusmaximus*) (Figueras *et al*., 2016), lenguado senegalés (*Solea senegalensis*) (Manchado *et al*., 2016) y hirame (*Paralichthysolivaceus*) (Shao*et al*., 2017). En este sentido, el disponer de genomas de otros peces planos permitirá realizar estudios de biología comparativa entre especies.

En este contexto, y con el objetivo a mediano termino de avanzar sobre el conocimiento de los mecanismos de diferenciación sexual y regulación del inicio de la pubertad en lenguado, nos hemos propuesto en el presente estudio secuenciar y ensamblar el genoma lenguado negro mediante la utilización de nuevas tecnologías de secuenciación masiva o *High throughputsequencing* a partir de distintas plataformas de secuenciación (NextGenerationSequencing-NGS).

MATERIALES Y MÉTODOS

1. *Obtención de muestra para secuenciar.*

En un muestreo rutinario de biometrías del stock de reproductores de lenguados (*Paralichthysorbignyanus*) del INIDEP se diseccionaron fragmentos de 1 x 1 cm de aleta caudal de 22 individuos salvajes. Las muestras fueron conservados en etanol al 96% en viales criogénicos de 1.8 mly posteriormente almacenados a 4°C. Los peces fueron anestesiados por inmersión en el MS-222 (Sigma-Aldrich). Todo el procedimiento de trabajo con peces se realizó bajo las normas del Comité Institucional para cuidado y uso de animales de experimentación (CICUAE) del IIB-INTECH.

2. *Extracción y comprobación de la cantidad y la calidad del ADN* genómico *(ADNg).*

La extracción del ADN se realizó en la Universidad de Santiago de Compostela, España utilizando el protocolo de NaCl. A fin de comprobar la calidad del ADN purificado, se llevó a cabo una electroforesis en gel de agarosa al 1% (**Figura 1**)**.**

COPIA ELECTRÓNICA INIDEP

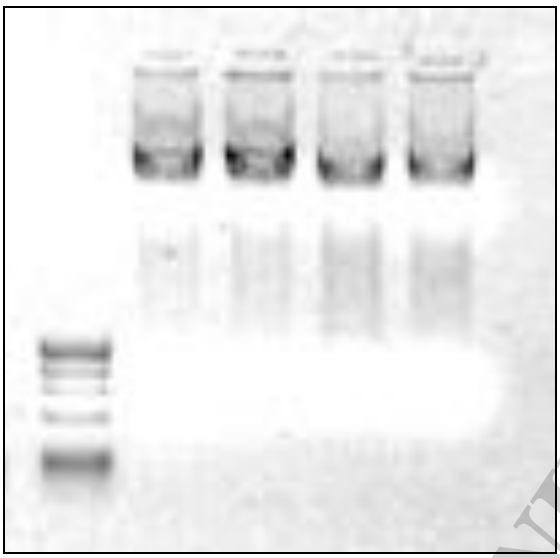

Figura 1. Foto del gel con las cuatro muestras de ADNg. Calle 1, Marcador de ADN, Calle 2: Hembra, Calle 3: Macho, Calle 4, *Pool* de Hembras, Calle 5: *Pool* de Machos.

Por último, las concentraciones se midieron en un espectrofotómetro (NanoDrop Technologies Inc.) y un fluorómetroQubit (Life Technologies) (**Tabla 1).**

Tabla 1. Control de calidad de las cuatro muestras de ADN por Nanodrop y Qubit.

Muestra	Nanod. ng/ul	A260	A280	260/280	260/230	Qubit. ng/ul
Hembra	48.41	0.968	0.526	1.84	2.02	41.8
Macho	57.99	1,160	0.631	1.84	1.99	49.8
Pool de Hembras	56.55	1,131	0.601	1.88	2.09	46.3
Pool de Machos	55.64	1,113	0.587	1.89	1.98	45.2

3. Secuenciación del genoma.

Para la secuenciación del ADNg recurrimos a los servicios de la empresa Oxford Genomics Centre (https://www.well.ox.ac.uk/ogc/) (Oxford, Inglaterra), que dispone de un secuenciador Illumina HiSeq-4000. Los fragmentos se secuenciaron por ambos extremos con lecturas de 150 nucleótidos de longitud (*paired-endreads*).

4. Ensamblado del genoma.

El análisis de los datos de la secuenciación masiva comenzaron a partir de los datos crudos (*raw data*) desde un formato de archivo FASTQ que fueron normalizados para obtener la información final. El ensamblado del organismo comenzó con; 1) Análisis de los *reads* (Fastqc). Para ver los requerimientos de filtrado y presencia de posibles

COPIA ELECTRÓNICA INIDEP

artefactos, filtrado y corte por calidad (con Prinseq y Trimomatic). 2) Corrección de los *reads*, mediante análisis estadísticos (Quake), se probarán distintos ensamblados con diferentes parámetros (tamaño de Kmer, bibliotecas*input*, etc.), utilizando diversos programas según requerimiento (Soapdenovo, Abyss, etc.). 3) Se seleccionó el mejor ensamblado en base a estadísticas generales y especificas del organismo (Quast, Assemblaton y bases de datos de KOG y DEG y homólogos al organismo), 4) Filtrado de contaminación (Deconseq) y separación de genoma nuclear y mitocondria.

5. *Búsqueda y anotación de genes.*

Para la anotación de los genes se siguieron los siguientes pasos: 1) Predicción de ARNs (tRNAscan, RNAmmer, Infernal), 2) entrenamiento de predictores de genes (*pipeline* propio CEBAR), 3) predicción de genes *de novo* (GeneId y GeneMark), por homología (Exonerate en base a proteínas cercanas o ESTs según disponibilidad), 4) generación de set consenso, 5) finalmente se realizó una anotación funcional (hmmer/Pfam para dominios, rpsblast/priam, lipoP,SignalP, THMM, TargetP, blast/KOG y asignación de GO). Los genes de interés identificados serán anotados en cuanto a nombre, estructura del gen, traducción a proteína y otras características.

*Nota***:** En la **Tabla suplementaria 1** se indican definiciones de la terminología básica utilizada en este estudio y en la **Tabla suplementaria 2** se enumeran los Software utilizados para análisis de datos.

RESULTADOS

1. *Ensamblado de novo del genoma.*

Se construyeron cuatro bibliotecas (hembra, macho, *pool* de 10 hembras y *pool* de 10 machos) de extremos pareados -*pairedend*- con una longitud de lectura de 150 pb y un tamaño de fragmento promedio de 350 pb (**Tabla 2**).

COPIA ELECTRÓNICA INIDEP

Tabla 2. Secuenciado del genoma (post-filtrado).

Muestra	Biblioteca	Longitud de lectura	Tamaño del Inserto	Cobertura
Macho	*pair-end*	150	~350bp	35.7X
Hembra	*pair-end*	150	~350bp	38.7X

Pool Machos	*pair-end*	150	~350bp	36.1X
Pool Hembras	*pair-end*	150	~350bp	35.5X

Las lecturas se filtraron primero para eliminar los adaptadores y las lecturas de baja calidad usando Trimmomatic (Bolger*et al*., 2014) y PrinSeq (Schmieder y Edwards, 2011) y luego se corrigió el error usando Quake (Kelley*et al*., 2010). Después del filtrado, se obtuvo una cobertura de más de 35X (**Tabla 2**). Las bibliotecas se ensamblaron individualmente utilizando SOAPdenovo 2.04 (Xie*et al*., 2014) con un tamaño de kmer de 29, luego se completaron los gaps con GapFiller (Boetzer y Pirovano, 2012) y finalmente se realizó un *scaffolding* con SSPACE3 con los parámetros predeterminados (Boetzer*et al*., 2011). Las ensamblados finales fueron de alrededor de 550 mb (**Tabla 3**).

Tabla 3. Estadística del ensamblado del genoma del lenguado negro.

Ensamblado	Macho	Hembra
# *contigs*	65,765	63,308
Tamaño *Scaffold* N50 (pb)	16.747	17.509
Scaffold más largo (pb)	191795	194365
Longitud total de *Scaffold* (pb)	540288577	538587025
GC Content (%)	41.38	41.39
%N	0.14	0.13

2. *Anotado del genoma.*

En una primera instancia se enmascararon los genomas utilizando el programa Repeatmasker (Hancock, 2004) y usando las bases de datos de Repbase (Tempel, 2008) y DFam (Hubley*et al*., 2016). En una segunda etapa y utilizando el genoma enmascarado se buscaron moléculas de ARN no codificante: ARNt, ARN de transferencia, ARNr, ARN ribosómicos y ARNnc, ARN no codificantes, usando los programas tRNAScan-SE (Zou*et al*., 2015), RNAmmer (Lagesen*et al*., 2007) e Infernal (Nawrocki*et al*., 2009) y la base de datos RFAM (Griffiths-Jones, 2003), respectivamente. En la tercera etapa se realizó una predicción de genes utilizando GeneId (Blanco y Abril, 2009) y Exonerate (Slater y Birney, 2005), entrenados con Uniprot / Swissprot (Pundir*et al*., 2017) y el genoma del hirame (*Paralichthysolivaceus*) (Shao*et al*., 2017). Por último, todas las proteínas se anotaron usando InterProScan (Hancock y Bishop, 2004). Para evaluar la calidad de los

COPIA ELECTRÓNICA INIDEP

ensamblados, se realinearon las lecturas filtradas de cada biblioteca a ambos genomas y se verificó que la mayoría de las lecturas estén asignadas (**Tabla 4**).

Tabla 4. Realineamiento de lecturas.

	Macho (*% alineado*)	Hembra(*% alineado*)
Hembra	*99,5*	-
Pool de machos	*99,9*	*99,4*
Pool de hembras	*99,5*	*99,9*

Para verificar la integridad de la anotación del genoma se verificaron los ortólogos por KOG (Tatusov*et al*., 2003) y BUSCO (Simão*et al*., 2015) (**Tabla 5**).

Tabla 5. Resultados de la anotación del genoma. BUSCO completos [C], BUSCO fragmentados (F), BUSCO faltantes (M). Total de grupos BUSCO buscados: 3640.

	Machos	Hembras
Proteínas	*25,193*	*25,231*
KOG	*458*	*458*
BUSCO	[C]: 37.9% [F]: 20.5% [M]: 41.6%	[C]: 37.9% [F]: 20.5% [M]: 41.6%
sin caracterizar	~4%	~4%
ARNt	474 (todos los tipos)	466 (todos los tipos)
ARNr	*15*	*15*
Otros ARNnc	*1432*	*1435*

DISCUSIÓN

La secuenciación completa del genomadel lenguado negro es la primera especie de un pez marino secuenciado hasta la actualidad en Argentina. El genoma del lenguado negro ha resultado ser muy parecido en tamaño y contenido de genes al genoma delhirame (Shao*et al*., 2017), especie del mismo género que el lenguado negro.La secuenciación del genoma utilizando tecnologías de NSG tienen muchas ventajas en comparación con otras metodologías de clonación de genes de la forma convencional. Dentro de sus principales ventajas encontramos que disponer del genoma del lenguado negro, una especie donde existen muy pocos estudios genéticos y secuencias de genes disponibles, permite no tener que clonar genes de forma individual.

COPIA ELECTRÓNICA INIDEP

La anotación del genoma, comprende la tarea de identificar y asignar funciones a los distintos elementos presentes en la secuencia genética de un organismo (Stein, 2001). Esta anotación puede realizarse de forma automática o manual. En el presente trabajo se realizó la anotación automática, sin embargo, si bien la anotación manual es más laboriosa para el estudio de genes de interés es recomendable valerse de ella a futuro por ser, en muchos casos, más precisa. El lograr caracterizar genes *in silico* dentro del genoma abre un importante camino para futuros estudios que busquen comprender y dilucidar los mecanismos moleculares y bioquímicos que regulan el eje reproductivo durante el proceso de determinación/diferenciación sexual, el inicio de la pubertad y la madurez sexual en el lenguado argentino, así como otros estudios que ayuden a mejorar el cultivo. Por otro lado, es importante remarcar que la información generada en este estudio, brindará la posibilidad de realizar futuros estudios de mejoramiento genético en el lenguado negro.

COPIA ELECTRÓNICA INIDEP

BIBLIOGRAFÍA

BAMBILL, G.A.; MASAKAZU, O.; RADONIC, M.; LÓPEZ, A.V.; MÜLLER, M.I.; BOCCANFUSO, J.J.; BIANCA, F.A. 2006. Broodstock management and induced spawning of flounder (*Paralichthysorbignyanus*) (Valenciennes, 1893) under a closed recirculated system. *Rev Biol Mar Oceanogr*, 41, 45-55.

BLANCO, E.; ABRIL, J.F. 2009.Computational Gene Annotation in New Genome Assemblies Using GeneID.*Methods MolBiol*, 537, 243–261.

BOCCANFUSO, J.J; ARISTIZABAL, A.E.O.; BERRUETA, M. 2019.Improvement of natural spawning of black flounder, *Paralichthysorbignyanus* (Valenciennes, 1839) by photothermal and salinity conditioning in recirculating aquaculture system.*Aquaculture*, 502, 134–141.

BOETZER, M.; HENKEL, C.V.; JANSEN, H.J.; BUTLER, D.; PIROVANO, W. 2011.Scaffolding pre‐assembled contigs using SSPACE.*Bioinformatics*, 27, 578–579.

BOETZER, M.; PIROVANO, W. 2012. Toward Almost Closed Genomes with GapFiller. *Genome Biol*, 13: R56.

BOLGER, A.M.; LOHSE, M.; USADEL, B. 2014.Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics*, 30, 2114–2120.

CHEN, S.; ZHANG, G.; SHAO, C.; HUANG, Q.; LIU, G.; ZHANG, P.; SONG, W.; AN, N.; CHALOPIN, D.; VOLFF, J.N.; HONG, Y.; LI, Q.; SHA, Z.; ZHOU, H.; XIE, M.; YU, Q.; LIU, Y.; XIANG, H.; WANG, N.; WU, K.; YANG, C.; ZHOU, Q.; LIAO, X.; YANG, L.; HU, Q.; ZHANG, J.; MENG, L.; JIN, L.; TIAN, Y.; LIAN, J.; YANG, J.; MIAO, G.; LIU, S.; LIANG, Z.; YAN, F.; LI, Y.; SUN, B.; ZHANG, H.; ZHANG, J.; ZHU, Y.; DU, M.; ZHAO, Y.; SCHARTL, M.; TANG, Q.; WANG, J. 2014. Whole-genome

sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *NatGenet*, 46, 253–260.

DÍAZ DE ASTARLOA, J.D.; MUNROE, T.A. 1998. Systematics, distribution and ecology of commercially important paralichthyd flounders occuring in Argentina-Uruguayan waters (*Paralichthys*, Paralichthydae). Anoverview. *J Sea Res*, 39, 1–9.

DÍAZ DE ASTARLOA, J.M. 2002. A review of the flatfish fisheries of the south Atlantic Ocean. *Rev Biol Mar Oceanogr*, 37, 113-125.

FIGUERAS, A.; ROBLEDO D.; CORVELO A.; HERMIDA M.; PEREIRO P.; RUBIOLO, J.A; GÓMEZ-GARRIDO, J.; CARRETÉ, L.; BELLO, X.; GUT, M.; GUT, I.G.; MARCET-HOUBEN, M.; FORN-CUNÍ, G.; GALÁN, B.; GARCÍA, J.L.; ABAL-FABEIRO, J.L.; PARDO, B.G.; TABOADA, X.; FERNÁNDEZ, C.; VLASOVA, A.; HERMOSO-PULIDO, A.; GUIGÓ, R.; ÁLVAREZ-DIOS, J.A.; GÓMEZ-TATO, A.; VIÑAS, A.; MASIDE, X.; GABALDÓN, T.; NOVOA, B.; BOUZA, C.; ALIOTO, T.; MARTÍNEZ, P. 2016. Whole Genome Sequencing of Turbot (*Scophthalmusmaximus*; Pleuronectiformes): A Fish Adapted to Demersal Life. *DNA Res*, 1-12.

GRIFFITHS-JONES, S.; BATEMAN, A.; MARSHALL, M.; KHANNA, A.; EDDY, S.R. 2003.Rfam: an RNA family database. *Nucleic Acids Res*, 31, 439–441.

HAIMOVICI, M.; MARTINS, A.S.; VIEIRA, P.C. 1996.Distribucao e abundancia de teléosteosdemersais sobre a plataforma continental do sul do Brasil. *Rev BrasBiol*, 56, 27–50.

HANCOCK, J.M. 2004.Repeatmasker*. In: Dictionary of Bioinformatics and Computational Biology*. John Wiley & Sons, Ltd, Chichester, p. dob0206.

HANCOCK, J.M.; BISHOP, M.J. 2004.InterProScan*. In: Dictionary of Bioinformatics and Computational Biology*. John Wiley & Sons, Ltd, Chichester, p.206.

HUBLEY, R.; FINN, R.D.; CLEMENTS, J.; EDDY, S.R.; JONES, T.A.; BAO, W.; SMIT, A.F.A.; WHEELER, T.J. 2016. The Dfam database of repetitive DNA families.*Nucleic Acids Res*, 44, 81–89.

KELLEY, D.R.; SCHATZ, M.C.; SALZBERG, S.L. 2010. Quake: Quality-Aware Detection and Correction of Sequencing Errors. *Genome Biol*, 11, R116.

LAGESEN, K.; HALLIN, P.; RØDLAND, E.A.; STAERFELDT, H.H.; ROGNES, T.; USSERY, D.W. 2007.RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*, 35, 3100-3108.

LÓPEZ, A.V.; RADONIC, M.; MÜLLER, M.I.; BAMBILL, G.A. 2009a.Influencia de la temperatura en el crecimiento y proporción de sexos en juveniles de lenguado (*Paralichthysorbignyanus*) nacidos en cautiverio. *RevAquaTIC*, 31, 1-5.

LÓPEZ, A.V.; MÜLLER, M.I.; RADONIC, M.; BAMBILL, G.A. BOCCANFUSO, J.J.; BIANCA, F.A. 2009b. Larval culture technique and quality control in juveniles of flounder*Paralichthysorbignyanus* (Valenciennes, 1839) in Argentina. *Spanish J Agric Res*, 7, 75-82.

MANCHADO, M.; PLANAS, J.V.; COUSIN, X.; REBORDINOS, L.; CLAROS, M.G. 2016. Current status in other finfish species: Description of current genomic resources for the gilthead seabream (*Sparusaurata*) and soles (*Soleasenegalensis* and *Soleasolea*). *Genomics in Aquaculture*, eds S. Mackenzie and S. Jentoft (Amsterdam: Elsevier), 195–221.

COPIA ELECTRÓNICA INIDEP

NAWROCKI, E.P.; KOLBE, D.L.; EDDY, S.R. 2009. Infernal 1.0: Inference of RNA Alignments. *Bioinformatics*, 25, 1335–1337.

PUNDIR, S.; MARTIN, M.J.; O'DONOVAN, C. 2017.UniProt Protein Knowledgebase.*Methods MolBiol*, 1558, 41–55.

RADONIC, M.; MÜLLER, M.I.; LÓPEZ, A.V.; BAMBILL, G.A.; SPINEDI, M.; BOCCANFUSO, J.J. 2007.Avances en la reproducción controlada del lenguado *Paralichthysorbignyanus* en Argentina. *Cienc Mar*, 33, 187-196.

RADONIC, M.; MACCHI, G.J. 2009.Gonadal sex differentiation in cultured juvenile flounder, *Paralichthysorbignyanus* (Valenciennes, 1839).*J World AquacultSocJ*, 40, 129-133.

TEMPEL, S.; JURKA, M.; JURKA, J. 2008.VisualRepbase: an interface for the study of occurrences of transposable element families. *BMC Bioinformatics*, 9, 345.

SCHMIEDER, R.; EDWARDS, R. 2011.Quality Control and Preprocessing of Metagenomic Datasets.*Bioinformatics*, 27, 863–64.

SHAO, C.; BAO, B.; XIE, Z.; CHEN, X.; LI, B.; JIA, X.; YAO, Q.; ORTÍ, G. 2017. The genome and transcriptome of Japanese flounder provide insights into flatfish asymmetry. *Nat Genet*, 49, 119-124.

SIMÃO, F.A.; WATERHOUSE, R.M.; IOANNIDIS, P.; KRIVENTSEVA, E.V; ZDOBNOV, E.M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212.

SLATER, G.S.; BIRNEY, E. 2005.Automated generation of heuristics for biological sequence comparison.*BMC Bioinformatics*, 6, 31.

STEIN, L. 2001. Genome annotation: from sequence to biology. *Nat Rev Genet*, 2, 493-503.

TATUSOV, R.L.; FEDOROVA, N.D.; JACKSON, J.D.; JACOBS, A.R.; KIRYUTIN, B.; KOONIN, E.V.; KRYLOV, D.M.; MAZUMDER, R.; MEKHEDOV, S.L.; NIKOLSKAYA, A.N.; RAO, B.S.; SMIRNOV, S.; SVERDLOV, A.V.; VASUDEVAN, S.; WOLF, Y.I.; YIN, J.J.; NATALE, D.A. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41.

XIE, Y.; WU, G.; TANG, J.; LUO, R.; PATTERSON, J.; LIU, S.; HUANG, W.; HE, G.; GU, S.; LI, S.; ZHOU, X.; LAM, T.W.; LI, Y.; XU, X.; WONG, G.K.; WANG, J. 2014. SOAPdenovo-Trans: *de novo*transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30, 1660–1666.

ZOU, Q.; GUO, J.; JU, Y.; WU, M.; ZENG, X.; HONG. 2015. Improving tRNAscan-SE Annotation Results via Ensemble Classifiers. *Mol Inform*, 34, 761–770.

COPIA ELECTRÓNICA INIDEP

Tabla Suplementaria 1. Secuenciación masiva: terminología básica.

Término o sigla	Definición
ADN	Ácido desoxirribonucleico.
Anotación	Asignación de una función a un gen conocido.
CEBAR	Centro de Bioinformática Argentina.
Contig	Secuencia de ADN que procede de dos o más secuencias que se superponen en sus extremos y se pueden juntar en una sola secuencia no redundante.
De novoassembly	Ensamblado de un genoma basado únicamente en la información que contienen las *reads*, sin necesidad de comparación con un genoma de referencia.
Ensamblado	Proceso por el cual los fragmentos cortos del ADN secuenciados se juntan en fragmentos más grandes hasta reconstruir el genoma.
EST	Acrónimo en inglés: *ExpressedSequenceTag*.
GAP	Ausencia de una o más bases en una de las hebras del DNA duplex.
k-mers	Subcadenas de caracteres en las que se compone cada secuencia y que son utilizadas por distintos procesos bioinformáticos, como el ensamblado *de novo* basado en gráficos de Bruijn por SPAdes o la asignación taxonómica de Kraken.
Library	Es una biblioteca de ADN.
NGS	*Nextgenerationsequencing* es la tecnología de secuenciación masiva que surgió después de la de Sanger.
PCR	Reacción en cadena de la polimerasa.
Paired-endsequencing	Secuenciación de lectura pareada que consiste en la secuenciación de fragmentos del ADN en ambos sentidos 5′ y 3.
Pipeline	Es una serie de múltiples software que se combinan para llevar a cabo un análisis automático determinado de forma secuencial o en paralelo.
Read	Lecturas o secuencia del ADN continua obtenida de un secuenciador.
Sequencecoverage	Es la cobertura o la parte estimada del genoma que ha sido secuenciada.
Sequencedepth	Profundidad de cobertura es la media del número de veces que cada base de un genomasecuenciando tiene una *read* que alinea en esa posición.
Scaffold	Serie de *contigs* ordenados, pero no necesariamente conectados en una secuencia sin huecos.

COPIA ELECTRÓNICA INIDEP

Tabla Suplementaria 2. Software para análisis de datos de secuenciación masiva.

Acción	Programa	Sitio web
Calidad de secuencia	*FastQC*	www.bioinformatics.babraham.ac.uk
	TRIMMOMATIC	http://www.usadellab.org/cms/?page=trimmomatic
	PRINSEQ	http://prinseq.sourceforge.net/
	DECONSEQ	http://deconseq.sourceforge.net
Identificación	*K-merFinder*	www.genomicepidemiology.org
	BLAST	www.ncbi.nlm.nih.gov/blast
	hmmer	http://hmmer.org
	PFAM	https://pfam.xfam.org
	rpsblast	http://nebc.nox.ac.uk/bioinformatics/docs/rpsblast.html
	PRIAM	http://priam.prabi.fr
Alineamiento	*lipoP*	http://www.cbs.dtu.dk/services/LipoP/
	SignalP	http://www.cbs.dtu.dk/services/SignalP/
	THMM	http://www.cbs.dtu.dk/services/TMHMM/
	TargetP	http://www.cbs.dtu.dk/services/TargetP/
	BLAST	www.ncbi.nlm.nih.gov/blast
	QUAKE	http://www.cbcb.umd.edu/software/quake/
	SOAPDENOVO	http://soap.genomics.org.cn/soapdenovo.html
Ensamblado	*ABYSS*	https://www.bcgsc.ca/resources/software/abyss
	QUAST	http://quast.sourceforge.net
	ASSEMBLATHON	https://assemblathon.org
	GapFiller	https://sourceforge.net/projects/gapfiller/
	SSPACE3	http://www.baseclear.com
	Repeatmasker	http://www.repeatmasker.org
	Repbase	https://www.girinst.org/repbase/
	DFam	https://dfam.org/home
	tRNAScan-SE	http://lowelab.ucsc.edu/tRNAscan-SE/
	RNAmmer	https://services.healthtech.dtu.dk/
	INFERNAL	http://eddylab.org/infernal/
	GENEID	https://genome.crg.cat/software/geneid/
	GENEMARK	http://exon.gatech.edu/GeneMark/
Anotado	*EXONERATE*	https://www.ebi.ac.uk/about/vertebrate-
	UNIPROT	https://www.uniprot.org
	SWISSPROT	https://web.expasy.org/docs/swiss-prot_guideline.html
	InterProScan	http://www.ebi.ac.uk/interpro/search/sequence/
	KOG	https://mycocosm.jgi.doe.gov/Tutorial/tutorial/kog.html
	DEG	http://tubic.tju.edu.cn/deg/
	BUSCO	https://busco.ezlab.org

COPIA ELECTRÓNICA INIDEP