eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
ZIENTZIA ETA TEKNOLOGIA FAKULTATEA
FACULTAD DE CIENCIA Y TECNOLOGÍA

Trabajo Fin de Grado
Grado en Biología

Caracterización genómica de la anchoa europea (*Engraulis encrasicolus*)

Autor/a:
Nerea Rubio Almeida
Director/a:
Luis Javier Chueca Simón
Codirector/a:
Iratxe Zarraonaindia Martínez

Leioa 21 de junio de 2023

Índice

Resumen

La anchoa europea, *Engraulis encrasicolus* (Linnaeus, 1758), es un pequeño pez teleósteo con un amplio rango de distribución que comprende la costa Atlántica de Europa y África occidental, el Mar Mediterráneo y el Mar Negro. Muestra una gran capacidad de dispersión, dando lugar a dos ecotipos genéticamente diferenciados como resultado de aislamientos, dispersiones y colonizaciones pasadas. Se trata de una especie sobreexplotada con gran importancia comercial, particularmente en el mar Cantábrico. La ausencia de recursos genómicos de la anchoa, su amplia distribución y dinámica poblacional, dificulta comprender sus requisitos fisiológicos y ecológicos. En este trabajo, se emplearon las tecnologías de secuenciación de Illumina y PacBio, para generar lecturas cortas (RNA-seq) y largas (HiFi) que fueron utilizadas para el posterior ensamblaje del transcriptoma y genoma, respectivamente, mediante herramientas bioinformáticas. Se obtuvo un genoma con una longitud de 918,56 megabases, formado por 27.218 cóntigos, una longitud de cóntigo N50 de 38,21 kilobases y con una integridad de BUSCO del 60%. Un 45,86% del ensamblaje está formado por elemento repetitivos y se predijeron un total de 35.742 genes codificantes de proteínas. Para complementar la anotación funcional del genoma se ensambló el transcriptoma de huevo de anchoa resultando en 349 megabases con una integridad del 48,7%. Además, se obtuvo el mitogenoma con una longitud de 16.677 pares de bases. Este estudio proporciona un conjunto de datos ómicos y el primer borrador del genoma para *Engraulis encrasicolus*, que aportará la base para la creación uno de mayor calidad. La disponibilidad de un genoma de referencia permitirá comprender su estructura y función, así como ser utilizado en estudios genómicos comparativos con otros cupléidos. A su vez, representa un recurso esencial para la conservación, gestión y explotación sostenible de la anchoa.

Abstract

The European anchovy, *Engraulis encrasicolus* (Linnaeus, 1758), is a small teleost fish with a wide distribution range that includes the Atlantic coast of Europe and western Africa, the Mediterranean Sea and the Black Sea. It has a great dispersal capacity, giving rise to two genetically differentiated ecotypes as a result of past isolations, dispersals, and colonizations. It is an overexploited species with great commercial importance, particularly in the Cantabrian Sea. The absence of genomic resources of the anchovy, its wide distribution and population dynamics, makes more difficult to understand its physiological and ecological requirements. In this work, Illumina and PacBio sequencing technologies were used to obtain short (RNA-seq) and long (HiFi) reads that were used for the subsequent assembly of the transcriptome and genome, respectively, using bioinformatics tools. A genome with a length of 918.56 megabases

was obtained, made up of 27,218 contigs, an N50 contig length of 38.21 kilobases and with a 60% BUSCO integrity. A 45.86% of the assembly was formed by repetitive elements and a total of 35,742 protein-coding genes were predicted. To complement the functional annotation of the genome, the anchovy egg transcriptome was assembled, resulting in 349 megabases with 48.7%. integrity. Furthermore, the mitogenome with a length of 16,677 base pairs was obtained. This study provides a set of omic data and the first draft genome of *Engraulis encrasicolus*, which will provide the basis for the creation of a higher quality one. The availability of a reference genome will allow to understand its structure and function, as well as to be used in comparative genomic studies with other clupleid fishes. In turn, it represents an essential resource for the conservation, management and sustainable exploitation of the anchovy.

1. Introducción

La anchoa europea, *Engraulis encrasicolus* (Linnaeus, 1758), es un pequeño pez teleósteo con un amplio rango de distribución costera y en la plataforma continental (Montes *et al*., 2016). Su extensión geográfica comprende la costa atlántica de Europa y África occidental, el Mar Mediterráneo y el Mar Negro (Ferrer *et al.*, 2016). Al igual que otras especies pelágicas, la anchoa muestra un comportamiento de cardumen y migratorio, así como una gran capacidad de dispersión tanto en estado larvario como adulto (Agostini y Bakun, 2002). Actualmente, se distinguen dos linajes con diferente comportamiento y morfología como resultado de aislamientos, dispersiones y colonizaciones pasadas (Zarraonaindia *et al.*, 2012). Se ha asociado esta diferenciación con la heterogeneidad del hábitat (Bembo *et al*., 1996), el clima (Silva *et al*., 2014), la distancia geográfica y las características oceánicas (Borrell *et al*., 2012; Zarraonaindia *et al*., 2012), que daría lugar a la existencia de dos ecotipos genéticamente diferenciados (Le Moan *et al.*, 2016).

Se trata de una especie con gran importancia comercial y pesquera internacionalmente, particularmente en el mar Cantábrico (Ferrer *et al*., 2016). En los últimos años, la alta demanda comercial ha desencadenado una sobreexplotación del recurso (Ferrer *et al*., 2016), donde históricamente numerosas pesquerías de anchoa han sufrido colapsos en sus stocks como en el mar Negro (1990), mar de Alborán (2001) y por último en el golfo de Bizkaia (2005), que no se recuperó hasta 2011. Debido a ello, y a pesar de que su abundancia sigue siendo alta, su biomasa y el tamaño medio de los individuos ha disminuido drásticamente (Van Beveren *et al*., 2014). De esta manera, la pesca es uno de los factores, que junto otros, influyen en la pérdida de diversidad génica (Ruggeri *et al*., 2016).

Su amplia distribución y dinámica poblacional, dificulta el seguimiento de su historia demográfica y la identificación de la procedencia de adultos, juveniles y huevos (Catanese *et al.*, 2020). Para comprender completamente la base genética desde el punto de vista evolutivo y ecológico, es necesario conocer la composición de genes y elementos reguladores de diferentes individuos o poblaciones. Para ello, se precisa conocer el genoma de *E. encrasicolus* como herramienta de referencia. De esta manera, se puede evaluar la estructura genética y dinámica de la(s) población(es) de anchoa. Así como, realizar estudios genéticos sobre su historia de vida y rasgos ecológicos, que será fundamental para la conservación de la especie y la correcta gestión del stock en la pesca.

El avance de las tecnologías de secuenciación de nueva generación ha supuesto una gran revolución en la obtención de genomas de organismos no modelo. Estas nuevas técnicas de secuenciación permiten obtener de manera masiva y paralela gran cantidad de secuencias de ADN y ARN en poco tiempo y con una mayor calidad y precisión. Actualmente, existen dos plataformas de secuenciación principales. Por un lado, las plataformas de secuenciación de lecturas cortas (secuenciación de segunda generación), representadas por *Illumina*, que producen lecturas de longitudes menores de 500 pares de bases (Alkan *et al*., 2010). Por otro lado, las plataformas de secuenciación de lecturas largas (secuenciación de tercera generación), representadas por *PacBio* y *Oxford Nanopore Technologies* (ONT), con una contigüidad y un perfil genético más completo, pero están limitadas debido a el rendimiento, el costo y la precisión (Metzker, 2010). La principal desventaja de la secuenciación de lecturas largas es su elevada tasa de error del 11-15 % (Rhoads & Au, 2015). Sin embargo, las lecturas individuales se pueden corregir usando, por ejemplo, lecturas de Illumina más precisas, que presentan una tasa de error de 1-1,5% (Cao *et al*., 2017). Alternativamente, se puede deducir una secuencia de consenso confiable a partir de lecturas de PacBio de alta cobertura (mayor de cincuenta veces) a través de herramientas bioinformáticas (Xie *et al*., 2020). Los sistemas PacBio más recientes muestran longitudes de lectura promedio muy mejoradas de más de diez kilobases, que permiten resolver regiones repetidas genómicas largas que no se pueden descifrar usando lecturas cortas de Illumina (Petit *et al*., 2017). Para mejorar la calidad de los ensamblajes, se desarrollaron lecturas de alta fidelidad (HiFi) para generar secuencias largas (con más de diez kilobases de longitud) con una tasa de error del 0,1% (Wenger *et al*., 2019).

La secuenciación PacBio, secuenciación de tercera generación (TGS), es un método de secuenciación a tiempo real y de forma continua, es decir, no hay pausa entre el proceso de

lectura. A diferencia de otros métodos de secuenciación, PacBio permite obtener longitudes de lectura mucho más largas y ejecuciones a mayor velocidad. El mecanismo se basa en ligar adaptadores de horquilla a ambos extremos de la secuencia de ADN objetivo, generando así un ADN circular monocatenario, denominado SMRTbell (Travers *et al*., 2010). Posteriormente, se carga la muestra en un chip (celda SMRT), con numerosos pocillos de unión "guía de ondas de modo cero" (ZMW). En la parte inferior de cada ZMW, se inmoviliza un ADN polimerasa encargado de la replicación (Rhoads & Au, 2015). Debido a que los nucleótidos se encuentran marcados con fluorescencia, a medida que la polimerasa retiene una base, se produce un pulso de luz que identificando la base unida (Eid *et al*., 2009). Después de que la polimerasa replica una hebra del ADN objetivo, como el ADN es monocatenario por la unión mediante los adaptadores, continúa incorporando las bases del adaptador y posteriormente la replicación de la hebra complementaria (Rhoads & Au, 2015). Este proceso se repite hasta que la polimerasa de degrada, de manera que se obtienen múltiples copias. Finalmente, se obtiene un registro de los pulsos de luz emitidos en cada ZMW, que permiten identificar la secuencia de bases (Rhoads & Au, 2015). A partir de las lecturas generadas, se puede obtener las lecturas HiFi (CCS), que derivan de una secuencia de consenso tras múltiples pases de un SMRTbell, produciendo lecturas más precisas a partir de sublecturas individuales ruidosas (Wenger *et al*., 2019).

El enfoque de Illumina logra una amplificación en puente de ADN, mediante la unión de fragmentos de ADN monocatenarios a una matriz (Steemers & Gunderson 2005). Los fragmentos de ADN a secuenciar se colocan sobre un soporte sólido, que contiene secuencias de ADN complementarias a los adaptadores de cada fragmento (Morozova, & Marra, 2008). De esta manera permite que cada fragmento se pueda anclar a la matriz (Morozova, & Marra, 2008). Una vez anclados los segmentos, la polimerasa los replica, generando una hebra reversa complementaria. La hebra original es retirada; mientras que la hebra reversa, a través de una secuencia terminal, se pliega y se ancla a su respectiva secuencia complementaria creando el "puente" (Morozova, & Marra, 2008). Posteriormente, la polimerasa genera una hebra complementaria idéntica a la original, que resulta en dos hebras clonadas del segmento inicial. Este proceso se repite masivamente hasta formar millones de copias de cada fragmento. Finalmente, se retiran las hebras reversas y se añaden oligonucleótidos marcados con fluorescencia de manera, que permite la identificación de la secuencia (Rubio *et al.*, 2020). Esto se repite de simultáneamente con todos los fragmentos. La exactitud de la secuenciación es determinada por la intensidad de la señal, y la longitud de las lecturas, por el número de ciclos realizados (Rubio *et al.*, 2020).

2. Objetivos

El objetivo de este estudio es generar por primera vez un borrador del genoma de la anchoa europea (siendo el segundo para la familia Engraulidae) que pueda servir como base para futuras investigaciones genómicas. Esto ayudará a ampliar el conocimiento sobre la familia Engraulidae y proporcionará una referencia para especies relacionadas. El ensamblaje del genoma permitirá analizar las dinámicas poblacionales, así como comprender la influencia de factores abióticos y bióticos en la especie y su diversidad genética.

3. Materiales y Métodos

3.1. Construcción de librerías y secuenciación

A partir de la extracción de ADN de un ejemplar adulto, se construyó una biblioteca SMRTbell siguiendo las instrucciones del kit "SMRTbell Express Prep v2.0 wiht Low DNA Input Protocol" (Pacific Biosciences, Menlo Park, CA). Se realizaron dos series de secuenciación de células SMRT en modo de secuenciación de consenso circular (CCS) en el Sequel System II con "Sequel II Sequencing Kit 2.0". También se estimó el tamaño del genoma mediante citometría de flujo.

Además, se secuenció el ARN total obtenido de 10 huevos y 3 larvas de anchoa mediante Illumina 150 ARN-seq de extremo emparejado, a partir de la construcción de dos librerías de ADN complementario de inserción de 250-300 pares de bases (pb), una para las muestras de huevos y otra para la de larvas. A partir de las lecturas de ADN y ARN obtenidas, se prosiguió con una serie de análisis bioinformáticos de los datos (Figura 1, Anexo I, Anexo II).

3.2. Ensamblaje del transcriptoma

Se comenzó analizando la calidad de las secuencias de ARN obtenidas a partir de las muestras de huevo y de larva mediante el programa FastQC v0.11.9 (Andrews, 2010) y se resumieron los resultados para su visualización utilizando la herramienta MultiQC v1.9 (Ewels *et al*., 2016). La secuencia de ARN de huevo se limpió con Trimommatic v0.39 (Bolger *et al*., 2014), que eliminó los adaptadores utilizados con Illumina y las secuencias de mala calidad (Phred < 30). A partir de estas secuencias, se ensambló el transcriptoma mediante Trinity v2.11.0 (Grabherr *et al*., 2011& Haas *et al*., 2013). La calidad del ensamblaje fue evaluada con BUSCO v4.1.4 (Manni *et al*., 2021) utilizando como base de datos de referencia "actinopterygii_odb10". Se analizó la distribución de las lecturas, mediante backmap v0.5 (Ewels *et al*., 2016; Li *et al*., 2009; Li, 2013; Li, 2018; Okonechnikov *et al*., 2016; Quinlan & Hall, 2010; R Core Team, 2021; Schell *et al.,* 2017) que alinea las lecturas de Illumina respecto al transcriptoma.

Para obtener una mayor representación de la expresión génica de la especie, se descargaron los datos de estudios previos disponibles en la base de datos *European Nucleotide Archive* (ENA). Del primer estudio (PRJNA348159) (ENA, 2016), se descargaron secuencias obtenidas a partir de juvenil y tejido de ovario, testículo, riñón, hígado y juvenil. El segundo estudio (PRJNA261165) (ENA, 2014), aportó secuencias obtenidas de tejido muscular. En ambos casos, tras eliminar los adaptadores mediante Trimommatic v0.39, se realizó una segunda limpieza de las secuencias mediante Cutadapt v2.8 (Martin, 2011) donde se eliminaron las primeras 15 bases de las secuencias para el estudio PRJNA348159 y las 10 primeras bases de las secuencias correspondientes al estudio PRJNA261165. Se comprobó los resultados con FastQC y MultiQC para continuar con el ensamblaje mediante Trinity v2.11.0, la evaluación de la calidad del ensamblaje con BUSCO v4.1.4 y el mapeo y cobertura mediante backmap v0.5. Los resultados obtenidos a partir de los tres análisis de la integridad del transcriptoma, mediante BUSCO, fueron combinados mediante el script de phyton3 "*generate_plot.py"*.

3.3. Ensamblaje del genoma

Las lecturas de ADN obtenidas mediante la secuenciación de PacBio se ensamblaron con dos herramientas diferentes, Hifiasm v0.16.1 (Cheng *et al*., 2021) y Flye v2.9 (Kolmogorov *et al*., 2019). La integridad de los ensamblajes obtenidos fue analizada utilizando BUSCO v5.2.2 con la base de datos odb10. Debido a que el genoma ensamblado con Hifiasm presentó la mayor contigüidad e integridad, fue seleccionado para análisis posteriores. Además, utilizando las lecturas obtenidas a partir de Hi-C, se realizó un tercer ensamblaje mediante Hifiasm v0.16.1 y su evaluación de integridad con BUSCO v5.2.2. Los dos genomas ensamblados con Hifiasm se compararon en términos de contigüidad usando Quast v5.0.2 (Gurevich *et al*., 2013).

Tras evaluar la comparativa entre ensamblajes, se decidió continuar con el genoma obtenido mediante Hifiasm y solo con las lecturas de PacBio. Para evaluar la distribución y cobertura de las lecturas frente al genoma, así como estimar el tamaño real del genoma, se empleó la herramienta backmap v0.5. Por último, se evaluó una posible contaminación en el ensamblaje usando BlobTools v1.1.1 **(**Laetsch & Blaxter, 2017), que evalúa la cobertura, el contenido de GC y la similitud de la secuencia contra cada secuencia de la base de datos NCBI-BLAST v2.12.0.

3.4. Anotación de repeticiones

Se ejecutó RepeatModeler v2.0 (Flynn *et al*., 2020) para construir *de novo*, a partir del ensamblaje, una biblioteca con las repeticiones. La biblioteca generada con las repeticiones de *Engraulis encrasicolus* se combinó con una biblioteca de repeticiones disponible para el pez

cebra (*Danio rerio*), (RepBase27.03.) (Bao *et al*., 2015; giriREPBASE, 2023), que se anotó y enmascaró usando RepeatMasker v4.1.4 (Tarailo & Chen, 2009). El genoma de referencia resultante, se alineó con las secuencias de ARN de diferentes tejidos (las de huevo obtenidas mediante Illumina y las procedentes de PRJNA348159) mediante HISAT2 (Kim *et al*., 2015).

3.5. Predicción de genes y anotación funcional

Después de mapear las secuencias repetidas, se realizó una predicción de genes por homología utilizando la herramienta GeMoMa v1.8 (Keilwagen *et al*., 2019) junto con 11 especies de la clase Actinopterygii como organismos de referencia (Anexo III, Anexo IV).

Primero, desde las lecturas mapeadas de ARN, los intrones se extrajeron y filtraron por los módulos GeMoMa ERE y DenoiseIntrons. Una vez eliminados los intrones, se ejecutó GeMoMa Pipeline para cada especie de referencia, como herramienta de alineación. Finalmente, las 11 anotaciones de genes se combinaron en una anotación final utilizando los módulos GeMoMa GAF y AnnotationFinalizer. Mediante GeMoMa Extractor, se obtuvieron tanto los CDS (secuencias codificantes) como las proteínas, las cuales fueron analizadas mediante BUSCO v5.2.2. A partir de las proteínas extraídas, InterProScan v5.39.77 (Jones *et al*., 2014) se utilizó para predecir motivos y dominios, así como su ontología génica (GO). Posteriormente, fueron anotados por búsqueda BLAST contra el Base de datos Uni-Prot con un límite de valor e de 10^{-6}.

A partir de la anotación funcional, se extrajeron las isoformas de las proteínas de mayor longitud y posteriormente se eliminó las duplicaciones con Agat v0.7.0 (Dainat, 2023). Después de cada anotación, se realizó una extracción de proteínas con GeMoMa Extractor para su posterior análisis con BUSCO v5.2.2.

3.6. Comparación del genoma ensamblado y anotación con especies relacionadas

Para la comparativa, mediante BUSCO v5.2.2, se utilizó el genoma y las proteínas extraídas de varias especies filogenéticamente cercanas (Anexo IV). Las especies a comparar fueron *Coilia nasus* (GCA_007927625.1, Xu *et al*., 2019), *Sardina pilchardus* (GCA_003604335.1, Louro *et al*., 2019) y *Alosa sapidissima* (GCA_018492685.1, Rhie *et al*., 2021), además de *E. encrasicolus*.

3.7. Genoma mitocondrial

Se ensambló el genoma mitocondrial a partir de las lecturas PacBio HiFi mediante la herramienta informática MitoHIFI (Uliano *et al*., 2023). Tras el ensamblaje, la anotación se

realizó con el programa MITOS v1 (Bernt *et al*., 2013) aplicando el código genético mitocondrial para vertebrados (tabla de traducción 2).

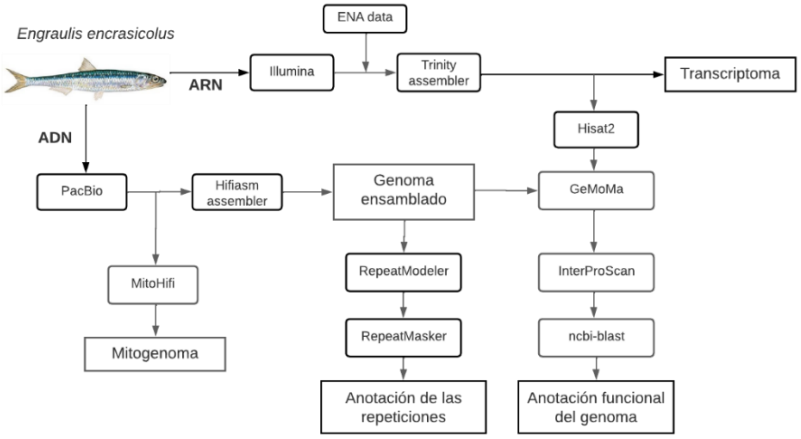


Figura 1. Representación gráfica del proceso seguido en este estudio para el ensamblaje del genoma.

4. Resultados

Tras la secuenciación se obtuvieron un total de 9,8 Gb en lecturas de Illumia ARN-Seq y 14 Gb de ADN (PacBio HiFi). El tamaño estimado del genoma mediante citometría de flujo fue de 1,74 Gb.

4.1.Transcriptoma

Debido a que las secuencias de ARN procedentes de la muestra de huevo eran de mayor calidad, estas fueron las utilizadas para la construcción del transcriptoma de 349Mb. Consultando el transcriptoma contra el conjunto de ortólogos del linaje "actinopterygii_odb10", se obtuvo que el 48,7% corresponden a genes completos (24,8 % genes de copia única y 23,9 % duplicados), el 14,1% a genes fragmentados y el 37,2% no se encontraron (Figura 2a). El conjunto de transcritos obtenidos a partir de ovario, testículo, riñón, hígado y juvenil (PRJNA348159), presentaron un mayor porcentaje de genes completos, 61,84%, de los cuales 30,52% son de copia única y el 31,31% duplicados (Figura 2a). Mientras que el 8,27% se encuentra fragmentado y 29,89% no encontrados. En cambio, para los transcritos procedentes del tejido muscular (PRJNA261165) solamente el 40,6% están completos (30,05% de copia única y 10,55% duplicados) y un 8,21% se encuentra fragmentado y 51,18% no encontrados (Figura 2a). Las 3 muestras presentaron 770 genes en común (Figura 2b). La muestra de huevo comparte 36 genes únicamente con la muestra de ovario, testículo, riñón, hígado y juvenil, mientras que 425 con la procedente de músculo. En contraposición, el músculo presentó 224 genes en común con la muestra de múltiples tejidos (Figura 2b).

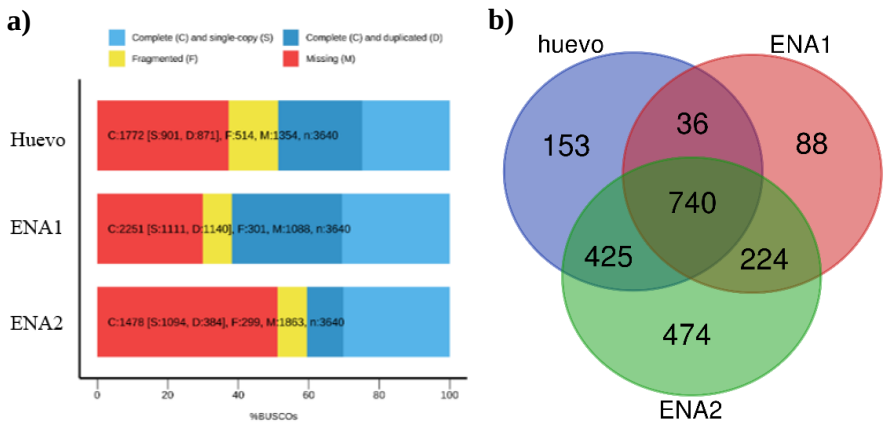


Figura 2. Comparativa de los transcriptomas procedentes de la muestra de huevo, la muestra de múltiples tejidos (ENA1) y la muestra de músculo (ENA2) de *E. encrasicolus*. **a)** Resultados de BUSCO y **b)** Diagrama de Venn con el número de genes expresados y la relación de estos entre las diferentes anotaciones.

4.2.Ensamblaje del genoma

La evaluación de integridad mediante BUSCO, mostró valores superiores para el ensamblaje mediante Hifiasm a partir de las secuencias procedentes de PacBio (Tabla 1). Presentó un 60,7% de genes completos (53,8% de copia única y 6,9% duplicados) en comparación con el 51,9% de Flye, de los cuales el 12,4% se encuentran duplicados y el 39,9% son de copia única. La proporción de genes completos recuperados fue menor en comparación con otras especies de actinopterigios, ya que se recuperó un 80,8%, 72,3% y 95,6% de genes completos para *Coilia nasus*, *Sardina pilchardus* y *Alosa sapidissima* respectivamente (Tabla 1).

Tabla 1. Estadísticos del ensamblaje del genoma de *Engraulis encrasicolus* y comparativa con otras especies filogenéticamente relacionadas.

	E. encrasicolus				***C. nasus***		***S. pilchardus***		***A. sapidissima***	
	Hifiasm		Flye							
	n	%	n	%	n	%	n	%	n	%
Completo:	2.208	60,7	1.890	51,9	2.942	80,8	2.632	72,3	3.480	95,6
Único	1.958	53,8	1.437	39,5	2.869	78,8	2.400	65,9	3.430	94,2
Duplicado	250	6,9	453	12,4	73	2	232	6,4	50	1,4
Fragmentado	271	7,4	215	5,9	148	4,1	257	7,1	62	1,7
No encontrado	1.161	31,9	1.535	42,2	550	15,1	751	20,6	98	2,7
Total	3.640		3.640		3.640		3.640		3.640	

El ensamblaje de las lecturas de PacBio comprendió 918,56 megabases (Mb) con longitudes de cóntigos N50 de 38,21 kilobases (kb) (Tabla 2). Más del 75% de la secuencia total estuvo

cubierta por 14.838 cóntigos. En cambio, los resultados para ensamblaje de Hi-C comprenden 932,03 Mb con longitudes de cóntigos N50 (Anexo V) de 37,24 kb (Tabla 2). Más del 75% de la secuencia total estuvo cubierta por 15.407 cóntigos. En ambos casos, el porcentaje de guanina y citosina fue del 43,6% y el cóntigo más largo abarcó 330,8 kb (Tabla 2).

Tabla 2. Estadísticas de Quast obtenidas del ensamblaje del genoma de *Engraulis encrasicolus*, a partir de las lecturas de PacBio y PacBio+Hi-C (PB-HiC).

Cóntigos	**PacBio**	**PB-HiC**
Número total	27.218	28.065
(>= 10.000 pb)	27.175	28.019
(>= 25.000 pb)	14.174	14.390
(>= 50.000 pb)	4.349	4.283
Mayor longitud	330.796	330.796
Longitud total:	918.567.494	932.031.078
(>= 10.000 pb)	918.155.782	931.593.577
(>= 25.000 pb)	672.596.428	674.180.618
(>= 50.000 pb)	335.753.269	328.670.160
N50	38.208	37.243
N75	24.247	23.907
L50	7.199	7.497
L75	14.838	15.407
GC (%)	43,6	43,6

El mapeo de las lecturas de PacBio contra el ensamblaje final del genoma mostró un total de 5,96 gigabases (Gb), con una cobertura máxima de 4 y una tasa de mapeo del 61,61% (Figura 3a). Por lo que el tamaño estimado del genoma, mediante backmap, es de 1,49 Gb, del cual el genoma obtenido representa un 61,65% (0,91857Gb).

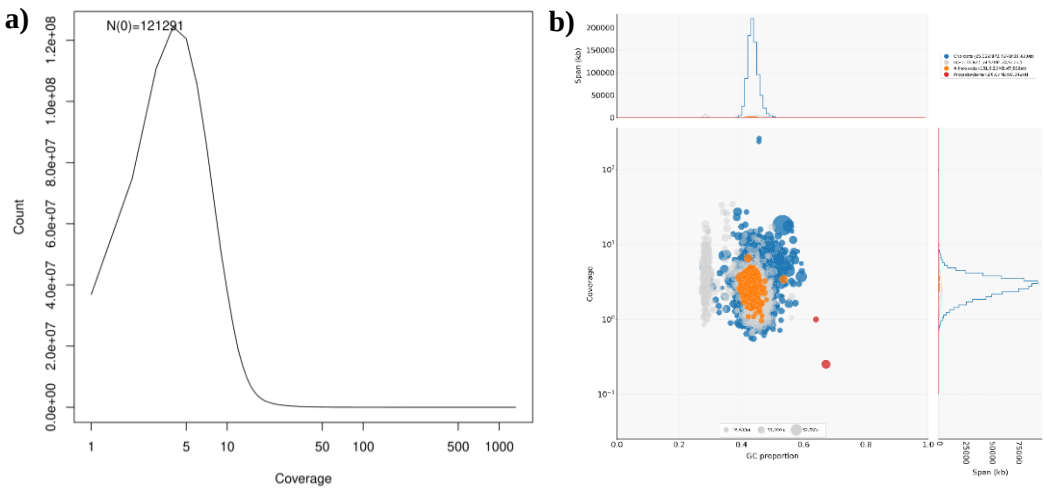


Figura 3. **a)** Distribución de cobertura por posición ensamblaje del genoma. **b)** Gráfico de burbujas que muestra la profundidad de lectura de cobertura. El tamaño de las burbujas

corresponde al tamaño de los cóntigos y los colores indican la anotación taxonómica correspondiente (gris=sin aciertos, azul=Chordata, naranja=Arthropoda y rojo=Proteobacteria). El tamaño de cada cóntigo y contenido de GC se proporciona en los paneles superior y derecho, respectivamente.

El análisis de BlobTools muestra una mínima fracción, del ensamblaje, asignada a distintos taxones (0,22% y 0,20% para Arthropoda y Proteobacteria respectivamente), en comparación con el 37,44% correspondiente a Chordata (Figura 3b).

4.3.Anotación de repeticiones y funcional del genoma

El contenido total de repeticiones del ensamblaje del genoma de *E. encrasicolus* se estimó en torno al 45,86 % (Tabla 3). Un 12,57% del genoma ensamblado se identificó como elementos transponibles, como elementos nucleares intercalados largos (LINE, 3,71 %), elementos nucleares intercalados cortos (SINE, 1,45 %), repeticiones en tándem largas (LTR, 7,41 %) y transposones de ADN (12,76%). El 20,52% correspondieron a familias de repeticiones no clasificadas. En menor proporción se distinguieron, repeticiones de ARN pequeño (1,19%), circular (0,74%) y otras repeticiones (satélite, 0,2%; repeticiones simples, 7,53% y de baja complejidad, 0,49%).

Tabla 3. Anotaciones repetidas basadas en homología y *de novo* según lo reportado por RepeatMasker y RepeatModeler.

	Número de elementos	**Longitud (pares de bases)**	**% de la secuencia**
Retroelementos	474.345	115.499.654	12,57
- SINES	83.457	13.335.649	1,45
- LINES	187.598	34.056.806	3,71
- LTR	203.290	68.107.199	7,41
Transposones ADN	582.285	117.238.957	12,76
Circular	27.271	6.804.689	0,74
No clasificadas	1.121.689	188.503.281	20,52
Total repeticiones		421.241.892	45,86
ARN pequeño	64.669	10.887.883	1,19
Satélite	2.499	1.854.931	0,2
Repeticiones simples	836.635	69.152.269	7,53
Baja complejidad	51.909	4.519.388	0,49

El genoma obtenido para la anchoa europea, tras enmascarar los elementos repetidos y eliminar las proteínas duplicadas, presentó una anotación con un total de total de 35.742 genes

funcionales, mismo número que de ARN mensajeros, con una longitud de 6.855 bases que abarcan, en conjunto, 325.944.113 bases del genoma (Tabla 4). Se estimó una presencia de 6 CDS por ARNm y 4.344 ARN mensajeros con un único CDS. Estos comprenden 221.433 secuencias codificantes de proteínas con una longitud media de 126 bases. En otras especies relacionadas, la longitud media de los CDS fue similar, sin embargo, el número de estos fue mayor en *A. sapidissima* y *Denticeps clupeoides.* En *S. pilchardus*, el número de genes y ARNm resultó mayor que en *E. encrasicolus* y *D. cupleoides*, pero con una longitud menor. A su vez la proporción de CDS por ARNm también fue menor en *S. pilchardus*, pero con un mayor número de secuencias únicas (Tabla 4).

Tabla 4. Estadísticos de anotación de los genes codificadores de proteínas predichos para el genoma de *Engraulis encrasicolus* y otras especies relacionadas (*Sardina pilchardus, Alosa sapidissima* y *Denticeps clupeoides*).

Especie	***E. encrasicolus***	***S. pilchardus***	***A. sapidissima***	***D. clupeoides***
Número:				
Gen	35.742	40.817	46.926	32.081
ARNm	35.742	40.847	56.007	46.831
CDS	221.433	201.277	718.000	586.377
Media:				
ARNms/gen	1	1,00073	2,17672	1,75627
CDSs/ARNm	6,19566	4,92758	12,7851	12,4681
Longitud media:				
Gen	6.855	2.292	3.578,5	6.236
ARNm	6.855	2.290	14.276	11.289
CDS	126	139	122	123
Espacio total:				
Gen	325.944.113	163.023.874	565.040.029	454.044.108
ARNm	325.944.113	163.023.874	535.671.349	443.806.982
CDS	47.773.982	43.936.487	47.741.741	46.489.192
Único:				
CDS ARNm	4.344	7.874	2.552	2.366

Tras la anotación, la evaluación mediante el análisis BUSCO (Tabla 5) se recuperó el 62% de genes codificantes de proteínas completos (54,8% único y 7,2% duplicado), con un 32% no encontrados y un 6% fragmentado. La proporción de genes completos fue similar al recuperado para las especies *C. nasus* y *S. pilchardus*, con 64,9% (62,7% único y 2,2% duplicado) y 62,9% (56,9% único y 6,1% duplicado) respectivamente. En cambio, para *A. sapidissima* se recuperó un 98,7% de genes completos (57% único y 41,7% duplicado), perdiendo un 0,9%.

Tabla 5. Estadísticos de la anotación funcional del genoma de *Engraulis encrasicolus* y otras especies relacionadas (*Coilia nasus*, *Sardina pilchardus y Alosa sapidissima*).

	E. encrasicolus		***C. nasus***		***S. pilchardus***		***A. sapidissima***	
	n	%	n	%	n	%	n	%
Completo:	2.256	62	2.362	64,9	2.289	62,9	3.595	98,7
Único	1.994	54,8	2.282	62,7	2.068	56,9	2.076	57
Duplicado	262	7,2	80	2,2	221	6,1	1.519	41,7
Fragmentado	21.900	6	216	5,9	365	10	16	0,4
No encontrado	1.165	32	1.062	29	986	27,1	29	0,9
Total	3.640		3.640		3.640		3.640	

4.4. Mitogenoma

El ensamblaje del mitogenoma, utilizando como referencia el construido por Lavoué *et al.* (2007), resultó en un cóntigo circularizado de 16.677pb de longitud (Figura 4). Su anotación reveló la recuperación de 37 genes mitocondriales, de los cuales 13 corresponden a genes codificadores de proteínas y 2 a ARN ribosómico y 22 a ARN de transferencia, siendo unos resultados similares a los presentados por Lavoue *et al*. (2007).

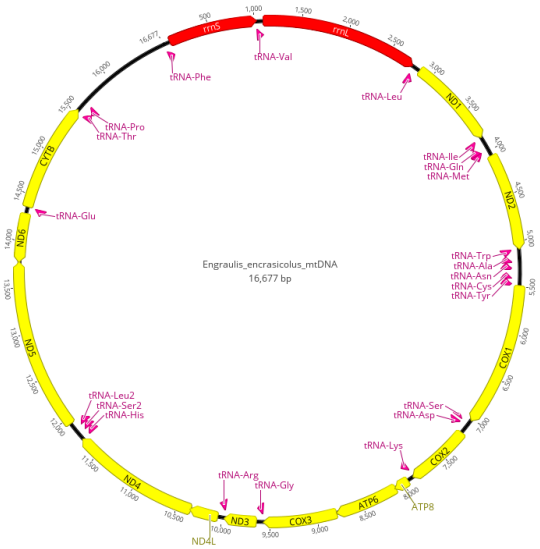


Figura 4. Genoma mitocondrial circular de *Engraulis encrasicolus*. Se proporcionan abreviaturas estándar para los genes de codificación de proteínas (amarillo), ARN de transferencia (rosa) y ribosómico (rojo). La orientación de los genes se indica mediante la dirección de las flechas.

5. Discusión

El transcriptoma ensamblado a partir de la muestra de huevo presentó una buena cobertura y alta calidad, recuperando prácticamente la mitad de los genes. Las diferencias en la

recuperación de genes completos dependen del origen de las lecturas, ya que la expresión génica varía en función del tejido y del momento. Los genes que se estén expresando en el huevo durante el desarrollo embrionario son diferentes a los que se expresan durante la fase adulta. Debido a ello, no se recupera la totalidad de los genes a partir del ARNm mediante BUSCO y los recuperados son diferentes para los transcriptomas ensamblados a partir de ARN-seq de distinto origen. La combinación de diferentes ensamblajes con datos de ARN-seq generados a partir de diferentes etapas de desarrollo ofrece un medio para producir transcriptomas de una calidad aún mayor, que a su vez darán como resultado un ensamblaje y anotación del genoma más completa.

El ensamblaje mediante Hifiasm y a partir de lecturas procedentes PacBio, ha sido la herramienta con mejores resultados, ya que recupera un mayor porcentaje de genes completos. Utilizando las lecturas de PacBio, el ensamblaje presenta mayor contigüidad, ya que se obtiene un N50 de 38,21 kb mientras que para Hi-C de 37,24 kb. En *C. nasus*, para un ensamblaje a nivel de cromosoma de 851,67 Mb, presenta un N50 de 35,42 Mb (Ma *et al*., 2023). Para un ensamblaje de mayor calidad se esperarían valores similares para *E. encrasicolus*, ya que una mayor longitud de N50 indica menor número de brechas en el ensamblaje y por tanto mayor contigüidad. La ventaja de haber utilizado secuencias de lectura largas, es que permite atravesar fácilmente las regiones más repetitivas y ayudar a llenar los espacios entre cóntigos, aumentando así la longitud de las secuencias ensambladas y, a su vez, mejora las estadísticas N50 (Logsdon *et al*., 2020). Se estima que el tamaño del genoma es de 1,49 Gb, menor a los esperado por citometría de flujo, con un 60% mapeado en el ensamblaje. Este coincide con el 60% de los genes recuperados en BUSCO, lo que parece indicar que se está perdiendo esa fracción del genoma. La pérdida de genes puede deberse a errores en la secuenciación, a una baja calidad de las lecturas secuenciadas o de la propia muestra, o a la base de datos utilizada para evaluar la integridad del ensamblaje del genoma. Utilizando como base de datos de referencia "actinopterygii_odb10" para el análisis BUSCO se recupera un 72,3% de genes completos para *S. pilchardus*, mientras que utilizando "actinopterygii_odb9" se recupera un 84,2 % (Louro *et al*.,2019). Esta es una versión anterior que contiene mayor número de genes ortólogos, que aumenta la posibilidad de recuperar mayor número de genes.

Las lecturas de PacBio utilizadas no presentan contaminación de otros organismos en la muestra. Blobtools no detectó contaminaciones según el contenido de GC y la distribución de cobertura. La detección mínima del filo Arthropoda o Proteobacteria probablemente se deba a secuencias altamente conservadas en los diferentes filos.

El contenido de repeticiones estimado para el genoma de la anchoa (45,48%) es ligeramente superior al porcentaje de elemento repetidos en *C. nasus* (41,32%) (Ma *et al.*, 2023) y en *S. pilchardus* (40,7 %) (Louro *et al*., 2019). La predicción del número genes codificantes de proteínas es menor para *E. encrasicolus*, que para *S. pilchardus* y *A. sapidissima*, ajustándose a lo esperado teniendo en cuenta que en el ensamblaje del genoma se recupera una mayor proporción de genes. A su vez, la integridad de los genes, codificantes de proteínas, anotados es del 62%, valor similar a la integridad del ensamblaje del genoma, lo que indica que para el ensamblaje obtenido la anotación es de alta calidad. Los valores de BUSCO son similares también para *C. nasus* (64,9%) y *S. pilchardus* (62,9%), mientras que para *A.sapidissima* recupera un 98,7%. Se observa que la integridad de los genes codificantes de proteínas es mayor cuanto mayor es la integridad del ensamblaje.

Por último, la longitud del genoma mitocondrial de *Engraulis encrasicolus* presenta una longitud similar (16.677 pb) a otras especies de la familia Engraulidae, como es *Engraulis rigens* con un mitogenoma de 16.690 pb de longitud (Sun, 2019). Esto sugiere que el ADN mitocondrial se encuentra muy conservado en esta familia e incluso en los vertebrados, ya que comparten la estructura con otros teleósteos y vertebrados (Chen *et al*., 2017; Shan *et al*., 2016; Wen *et al*., 2017; Zou *et al*., 2017).

6. Conclusión

En el presente estudio, se utilizó una estrategia combinada que involucra las tecnologías Illumina y PacBio para el ensamblaje *de novo* del genoma y transcriptoma del huevo de *Engraulis encrasicolus.* Se obtuvo un genoma con una longitud de 918,56 Mb, estimando que corresponde al 60% del tamaño total del genoma, formado por 27.218 cóntigos y una longitud de cóntigo N50 de 38,21 kb. Un 45,86% del ensamblaje está formado por elemento repetitivos y se predijeron un total de 35.742 genes codificantes de proteínas. Además, se construyó el mitogenoma que proporciona las bases para detectar marcadores mitocondriales de alta resolución que permiten la identificación de especies. El borrador del genoma obtenido por primera vez en este trabajo, así como la predicción de genes, y su anotación funcional, servirá como punto de partida para obtener un ensamblaje de mayor calidad, a nivel de cromosoma. Además, este genoma servirá como un recurso para futuros estudios genómicos, evolutivos, poblacionales y de la biología de la conservación de la anchoa europea. Para ello, es necesario secuenciar una mayor cantidad de lecturas largas con la finalidad de intentar mejorar la contigüidad e integridad del genoma obtenido.

7. **Bibliografía**

Agostini, V. N. & Bakun, A. (2002). 'Ocean triads' in the Mediterranean Sea: physical mechanisms potentially structuring reproductive habitat suitability (with example application to European anchovy, *Engraulis encrasicolus*). *Fisheries Oceanography, 11*(2), 129-142. doi:10.1046/j.1365-2419.2002.00201.x.

Alkan, C., Sajjadian, S., & Eichler, E. E. (2010). Limitations of next-generation genome sequence assembly. *Nature Methods, 8*, 61-65. doi: 10.1038/nmeth.1527.

Bao, W., Kojima, K. K. & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA, 6*, 11. doi: 10.1186/s13100-015-0041-9.

Bembo, D. G., Carvalho, G. R., Cingolani, N., Arneri, E., Giannetti, G. & Pitcher, T. J. (1996). Allozymic and morphometric evidence for two stocks of the European anchovy *Engraulis encrasicolus* in Adriatic waters. *Marine Biology, 126*, 529-538. doi:10.1007/BF00354635.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., … Stadler, P. F. (2013). MITOS: Improved de novo Metazoan Mitochondrial Genome Annotation. *Molecular Phylogenetics and Evolution, 69*(2), 313-319. doi:10.1016/j.ympev.2012.08.023.

Borrell, Y. J., Piñera, J. A., Prado, J. A. S. & Blanco, G. (2012). Mitochondrial DNA and microsatellite genetic differentiation in the European anchovy *Engraulis encrasicolus* L. *ICES Journal of Marine Science: Journal du Conseil, 69*(8), 1357-1371. doi:10.1093/icesjms/fss129.

Cao, Y., Fanning, S., Proos, S., Jordan, K. & Srikumar, S. (2017). A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology, 8*, 18-29. doi:10.3389/fmicb.2017.01829.

Catanese, G., Di Capua, I., Iriondo, M., Bonanno, A., Estonba, A. & Procaccini, G. (2020). Application of high-throughput single nucleotide polymorphism genotyping for assessing the origin of *Engraulis encrasicolus* eggs. *Aquatic Conservation*: *Marine and Freshwater Ecosystems*, *30*, 1313-1324. doi:10.1002/aqc.3321.

Chen, Z., Li, H., Zhu, Y., Feng, Q., He, Y. & Chen, X. (2017). Molecular phylogeny of the family Dicroglossidae (Amphibia: Anura) inferred from complete mitochondrial genomes. *Biochemical Systematics and Ecology, 71*, 1-9. doi:10.1016/j.bse.2017.01.006.

Cheng, H., Concepcion, G. T., Feng, X., Haowen, Z, & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods, 18*, 170-175. doi:10.1038/s41592-020-01056-5.

Dainat, J. (s.f.). AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. (Version v0.7.0). *Zenodo*. doi:10.5281/zenodo.3552717.

Eid, J., Fehr, A. Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science, 323*(5910), 133-138. doi:10.1126/science.1162986.

European Nucleotide Archive (ENA). 2013. *Genome-wide transcriptome profiling of anchovy muscle transcriptome.* Recuperado de European Nucleotide Archive el 6 de febrero de 2023.

European Nucleotide Archive (ENA). 2016. *Global tissue-spesific transcriptome analysis for Engraulis encrasicolus.* Recuperado de European Nucleotide Archive el 6 de febrero de 2023.

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. doi:10.1093/bioinformatics/btw354.

Ferrer, D. M., Lloret, J., Muñoz, M., Faliex, E., Vila, S. & Sasal, P. (2016). Links between parasitism, energy reserves and fecundity of European anchovy, *Engraulis encrasicolus*, in the northwestern Mediterranean Sea. *Conservation Physiology*, *4*(1), cov069. doi:10.1093/conphys/cov069.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451-9457. doi:10.1073/pnas.1921046117.

giriREPBASE. 2023. *RepBase27.03.* Recuperado de giriREPBASE el 5 de mayo de 2023.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. doi: 10.1093/bioinformatics/btt086.

Jones, P., Binns, D. Chang, H. Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236-1240. doi:10.1093/bioinformatics/btu031.

Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods in molecular biology (Clifton, N.J.)*, *1962*, 161-177. doi:10.1007/978-1-4939-9173-0_9.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, *12*(4), 357-360. doi:10.1038/nmeth.3317.

Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology, 37,* 540-546. doi:10.1038/s41587-019-0072-8.

Laetsch, D. R. & Blaxter M. L. (2017). BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. *F1000Research*, *6*, 1287. doi:10.12688/f1000research.12232.1.

Lavoué, S., Miya, M., Saitoh, K., Ishiguro, N. B & Nishida, M. (2007). Phylogenetic relationships among anchovies, sardines, herrings and their relatives (Clupeiformes), inferred from whole mitogenome sequences. *Molecular Phylogenetics and Evolution*, *43*(3), 1096-105. doi:10.1016/j.ympev.2006.09.018.

Le Moan, A., Gagnaire, P. A. & Bonhomme, F. (2016). Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, *25*(13), 3187-3202. doi:10.1111/mec.13627.

Li, C. & Ortí, G. (2007). Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, *44*(1), 386-398. doi:10.1016/j.ympev.2006.10.030.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Preprint arXiv, 1303.3997*. doi:10.48550/arXiv.1303.3997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100. doi:10.1093/bioinformatics/bty191.

Logsdon, G. A., Vollger, M. R. & Eichler, E. E. (2020). Secuenciación del genoma humano de lectura larga y sus aplicaciones. *Nature Reviews Genetics, 21*(10), 597-614. doi:10.1038/s41576-020-0236-x.

Louro, B., De Moro, G., Garcia, C., Cox, C. J., Veríssimo, A., Sabatino, S. J., … Canário, A. V. M. (2019). A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*). *GigaScience, 8*(5), giz059. doi:10.1093/gigascience/giz059.

Ma, F., Wang, Y., Su, B., Zhao, C., Yin, D., Chen, C., … Liu, K. (2023). Gap-free genome assembly of anadromous *Coilia nasus*. *Scientific Data, 10*, 360. doi:10.1038/s41597-023-02278-w.

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols, 1*, e323. doi: 10.1002/cpz1.323

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal, 17*(1), 10-12. doi: 10.14806/ej.17.1.200.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics, 11*, 31-46. doi: 10.1038/nrg2626.

Montes, I., Zarraonaindia, I., Iriondo, M., Stewart W. G., Manzano, C., Cotano, U., … Estonba, A. (2016). Transcriptome analysis deciphers evolutionary mechanisms underlying genetic differentiation between coastal and offshore anchovy populations in the Bay of Biscay. *Marine Biology, 163*, 205. doi: 10.1007/s00227-016-2979-7.

Morozova, O. & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics, 92*(5), 255-264. doi:10.1016/j.ygeno.2008.07.001.

Okonechnikov, K., Conesa, A. & García, F. (2016). Qualimap2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics, 32*(2), 292-294. doi:10.1093/bioinformatics/btv566.

Petit, J., David, L., Dirks, R. & Wiegertjes, G. F. (2017). Genomic and transcriptomic approaches to study immunology in cyprinids: What is next? *Developmental & Comparative Immunology, 75*, 48-62. doi:10.1016/j.dci.2017.02.022.

Quinlan, A. R & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033.

R Core Team. (2021). R*: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Consultado el 20 febrero de 2023 en http://www.R-project.org/

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature, 592*(7856), 737-746. doi:10.1038/s41586-021-03451-0.

Rhoads, A. & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics, 13*(5), 278-289. doi:10.1016/j.gpb.2015.08.002.

Rubio, S., Pacheco, R. A. O., Milena, A. M., Perdomo, S. & García, R. R. (2020). Secuenciación de nueva generación (NGS) de ADN: presente y futuro en la práctica clínica. *Universitas Medica*, *61*(2). doi:10.11144/Javeriana.umed61-2.sngs.

Ruggeri, P., Splendiani, A., Di Muri, C., Fioravanti, T., Santojanni, A., Leonori, I., … Caputo, V. B. (2016). Coupling Demographic and Genetic Variability from Archived Collections of European Anchovy (*Engraulis encrasicolus*). *PLOS ONE, 11*(3), e0151507. doi: 10.1371/journal.pone.0151507.

Shan, B., Song, N., Han, Z., Wang, J., Gao, T. & Yokogawa, K. (2016). Complete mitochondrial genomes of three sea basses Lateolabrax (Perciformes, Lateolabracidae) species: Genome description and phylogenetic considerations. *Biochemical Systematics and Ecology*, *67,* 44-52. doi:10.1016/j.bse.2016.04.007.

Silva, G., Lima, F. P., Martel, P. & Castilho, R. (2014). Thermal adaptation and clinal mitochondrial DNA variation of European anchovy. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1792), 20141093. doi:10.1098/rspb.2014.1093.

Steemers, F. J & Gunderson, K. L. (2005). Illumina, Inc. *Pharmacogenomics, 6*(7), 777-782. doi:10.2217/14622416.6.7.777.

Sun, W. (2019). The complete mitochondrial genome of *Engraulis ringens* (Engraulidae, Clupeiformes) and phylogenetic studies of Engraulidae. *Mitochondrial DNA Part B*, *4*(2), 3525-3526. doi: 10.1080/23802359.2019.1675553.

Tarailo, M. G., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, *25*(1), 4.10.1-4.10.14. doi:10.1002/0471250953.bi0410s25.

Travers, K., Chin, C. S., Rank, D., Eid, J. & Turner, S. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159. doi:10.1093/nar/gkq543.

Uliano, M. S., Ferreira, J. G., Krasheninnikova, K., Darwin Tree of Life Consortium, Formenti, G., Abueg, L., … McCarthy, S. A. (2023). MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads. *bioRxiv*. doi:10.1101/2022.12.23.521667

Van Beveren, E., Bonhommeau, S., Fromentin, J. M., Bigot, J. L., Bourdeix, J. H., Brosset, P., … Saraux, C. (2014). Rapid changes in growth, condition, size and age of small pelagic fish in the Mediterranean. *Marine Biology, 161,* 1809-1822. doi:10.1007/s00227-014-2463-1

Wen, Z. Y., Xie, B. W., Qin, C. J., Wang, J., Yuan, D. Y., Li, R. & Zou, Y. C. (2017). The complete mitochondrial genome of a threatened loach (*Beaufortia kweichowensis*) and its phylogeny. *Conservation Genetic Resources, 9,* 565-568. doi:10.1007/s12686-017-0723-3.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepción, G. T., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155-1162. doi:10.1038/s41587-019-0217-9.

Xie, H., Yang, C., Sun, Y., Igarashi, Y., Jin, T. & Luo, F. (2020). PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning. *Frontiers in Genetics*, *11,* 1664-8021. doi:10.3389/fgene.2020.516269.

Xu, G., Bian, C., Nie, Z., Li, J., Wang, Y., Xu, D., … Xu, P. (2019). Supporting data for "Genome and population sequencing of a chromosome-level genome assembly of Chinese tapertail anchovy (*Coilia nasus*) provides novel insights into migratory adaptation" *GigaScience Database*. doi:10.5524/100677.

Zarraonaindia, I., Iriondo, M., Albaina, A., Pardo, M. A., Manzano, C., Stewart, W. G., … Estonba, A. (2012). Multiple SNP markers reveal fine scale population and deep phylogeographic structure in European anchovy (*Engraulis encrasicolus* L.). *PLoS One*, *7*, e42201. doi:10.1371/journal.pone.0042201.

Zou, Y. C., Xie, B. W., Qin, C. J., Wang, Y. M., Yuan, D. Y., Li, R. &Wen, Z. Y. (2017). The complete mitochondrial genome of a threatened loach (*Sinibotia reevesae*) and its phylogeny. *Genes & Genomics*, *39*, 767- 778. doi:10.1007/s13258-017-0541-8.

8. Anexo e información complementaria

Anexo I. Software utilizados en este estudio, su versión y disponibilidad de fuente. Se accedió por última vez a todas las URL el 14-06-2023.

Nombre	Versión	Enlace
AGAT	0.7.0	https://github.com/NBISweden/AGAT
AUGUSTUS	3.4.0	https://github.com/Gaius-Augustus/Augustus
backmap	0.5	https://github.com/schellt/backmap
BlobTools	1.1.1	https://github.com/DRL/blobtools
BUSCO	4.1.4 & 5.2.2	https://busco.ezlab.org/
Cutadapt	2.8	https://cutadapt.readthedocs.io/en/stable/
FastQC	0.11.9	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Flye	2.9	https://github.com/fenderglass/Flye
GeMoMa	1.8	http://www.jstacs.de/index.php/GeMoMa
Hifiasm	0.16.1	https://github.com/chhylp123/hifiasm
HISAT2	2.2.1	http://daehwankimlab.github.io/hisat2/
Interproscan	5.39.77	https://github.com/ebi-pf-team/interproscan
Maker	2.31.10	https://github.com/Yandell-Lab/maker
MitoHIFI	2.2	https://github.com/marcelauliano/MitoHiFi
MITOS	1	http://mitos.bioinf.uni-leipzig.de/index.py
MultiQC	1.9 & 1.10	https://multiqc.info/
Ncbi-blast	2.14.0	https://blast.ncbi.nlm.nih.gov/Blast.cgi
RepeatMasker	4.1.4	https://www.repeatmasker.org/
RepeatModeler	2.0	https://github.com/Dfam-consortium/RepeatModeler/blob/master/RepeatModeler
Samtools	1.15.1	https://github.com/samtools/samtools
Trimmomatic	0.39	http://www.usadellab.org/cms/?page=trimmomatic
Trinity	2.11.0	https://github.com/trinityrnaseq/trinityrnaseq/wiki
Quast	5.0.2	https://github.com/ablab/quast

Anexo II. Disponibilidad de material suplementario.

El código empleado en este trabajo está subido en un repositorio de Github, que se puede consultar en: https://github.com/ljchueca/Engraulis_encrasicolus_genome.git.

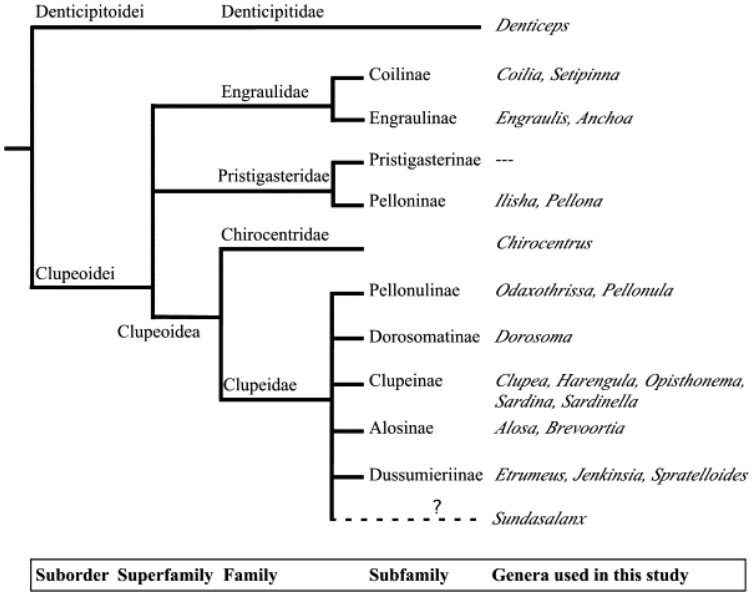
Anexo III. Listado de los organismos cuyo genoma y anotación se utilizó como referencia para la anotación funcional de *Engraulis encrasicolus.*

Orden	Familia	Especie	Acceso GenBank
Clupeiformes	Engraulidae	*Coilia nasus*	*
Clupeiformes	Clupeidae	*Alosa sapidissima*	GCA_018492685.1
Clupeiformes	Clupeidae	*Alosa alosa*	GCA_017589495.2
Clupeiformes	Clupeidae	*Clupea harengus*	GCA_900700415.2
Clupeiformes	Denticipitidae	*Denticeps clupeoides*	GCA_900700375.1
Cypriniformes	Cyprinidae	*Cyprinus carpio*	GCA_018340385.1
Cypriniformes	Cyprinidae	*Carassius auratus*	GCA_003368295.1
Carangiformes	Carangidae	*Seriola lalandi*	GCA_002814215.1

Orden	Familia	Especie	Acceso GenBank
Centrarchiformes	Centrarchidae	*Micropterus salmoides*	GCA_014851395.1
Batrachoidformes	Batrachoididae	*Thalassophryne amazonica*	GCA_902500255.1
Perciformes	Cyclopteridae	*Cyclopterus lumpus*	GCA_009769545.1

* El genoma de esta especie no está disponible en el GenBank, se ha obtenido de la base de datos GigaScience Database (Xu *et al*., 2019).

Anexo IV. Árbol filogenético del orden Clupeiformes basado en secuencias de ADN nuclear y mitocondrial (Li & Ortí, 2017).



Anexo V. Glosario con abreviaturas y terminología básica.

Término o sigla	Definición
ADN	Ácido desoxirribonucleico.
Anotación	Asignación de una función a un gen conocido.
ARN	Ácido ribonucleico.
ARN-seq	Secuenciación de ARN.
Biblioteca	Representación del conjunto de ADN o ADN complementario. La de ADNc se construye mediante retrotranscripción de los ARNm, y por lo tanto solo representa las regiones codificantes de proteínas del genoma.
CDS	Coding Sequence es parte del ARNm o secuencia genómica que codifica una secuencia de proteína.
Cobertura	Estimación de la proporción del genoma que ha sido secuenciada.
Cóntigo	Secuencia contigua: 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.
Ensamblaje d*e novo*	Ensamblado de un genoma basado únicamente en la información que contienen las lecturas, sin necesidad de comparación con un genoma de referencia.

Término o sigla	Definición
Ensamblado	Proceso por el cual los fragmentos cortos del ADN secuenciados se juntan en fragmentos más grandes hasta reconstruir el genoma.
Gb	Gigabase: unidad de medida para designar la longitud del ADN. Es igual a mil millones de bases.
Kb	Kilobase: unidad de medida para designar la longitud del ADN. Es igual a mil de bases.
Lectura	Secuencia del ADN continúa obtenida de un secuenciador.
L50	Dado un conjunto de cóntigos, cada uno con su propia longitud, el L50 se define como el recuento del menor número de cóntigos cuya longitud suma la mitad del tamaño del genoma.
L75	El recuento del menor número de cóntigos cuya longitud suma el 75% del tamaño del genoma.
Mapeado	Alineamiento de cada una de las lecturas de ADN a una posición en el genoma de referencia.
Mb	Megabase: unidad de medida para designar la longitud del ADN. Es igual a 1 millón de bases
NGS	Next generation sequencing es la tecnología de secuenciación masiva que surgió después de la de Sanger.
N50	Es la longitud máxima de cóntigo que representa al menos el 50 % de la longitud total del ensamblaje.
N75	Es la longitud máxima de cóntigo que representa al menos el 75 % de la longitud total del ensamblaje.
PCR	Reacción en cadena de la polimerasa.
Profundidad	Media del número de veces que cada base de un genoma secuenciando tiene una lectura que alinea en esa posición.