eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

DOCTORAL THESIS

Micro and Macro-Evolutionary Studies in Non-Model Species: a Transcriptomic Perspective in Teleosts

Author:
Jorge Eliseo Langa Arranz

Supervisors:
Dr. Andone Estonba Rekalde
Dr. Darrell Conklin

Department of Genetics, Physical Anthropology and Animal Physiology
Faculty of Science and Technology

2021

Summary

Genomic resources and bioinformatics tools are very scarce in non-model species, such as most fish species, yet there are sufficient ecological and economic reasons for this situation to change. RNA-Seq would be an effective tool for generating genomic resources aimed at revealing genetic variation and function necessary for both micro- and macro-evolutionary studies in non-model species. My Thesis is a natural continuation of previous works carried out in my research group. This research line beginned with a population genetics study of the European anchovy (*Engraulis encrasicolus, L.*) by Montes et al. (2013), which needed the development of the IEB method by Conklin et al. (2013). These studies were followed by a similar application of the methodology on Atlantic mackerel (*Scomber scombrus, L.*; Genomic Resources Development Consortium et al., 2015).

This Thesis characterizes several non-model fish species at two different time scales: at the population level in *Tinca tinca*, a freshwater fish native to rivers across Europe and Central Asia, and at the inter-species level in Clupeiformes, present worldwide both in salt and freshwater. Both studies have in their foundations RNA Sequencing. This Thesis also contributes with bioinformatic methods and pipelines for *de novo* transcriptome assembly, and the identification of (1) its sequences, (2) their functions, (3) exon structure, (4) nucleotide variation, and (5) genes under positive selection.

The first section of this Thesis, published in "A novel transcriptome-derived SNPs array for tench (Tinca tinca L.)" (Kumar et al., 2019), addresses the presence or absence of the Western and Eastern phylogroups in two cultured populations of *Tinca tinca* in Central Europe. To this end, genetic variation was ascertained within and between two cultured tench breeds from the Czech Republic and Hungary. This is the first study that generates genomic and transcriptomic resources for this species.

We sampled RNA from individuals from both sexes, two tissues (brain and muscle) and two different metabolic rates (fast in summer, and slow during winter). We assembled a transcriptome composed of 267,058 sequences, annotating it with functional information, and predicting the set of coding sequences and protein translations. With supplementary DNA sequencing of other ten individuals, we discovered 60,414 Single Nucleotide Polymorphisms (SNP). From them we constructed a 96 SNP chip for this species, the first one, and used it to genotype 140 samples

with pure and mixed phylogroup ancestries (Eastern and Western, according to the growth hormone gene), from two local breeds in the Czech Republic and Hungary.

Clustering results indicate that the most probable number of ancestries is two, the same as the number of phylogroups, but most individuals have a mixed ancestry. According to the F_{ST} statistic, there exist small but significant differences between breeds, but not at the phylogroup (Western, Eastern, or Hybrid) level. Therefore, within the breed there is gene flow between individuals of both phylogroups, and therefore there is no reproductive isolation between the two phylogroups. Our study supports the hypothesis that the analyzed individuals would turn out to be a genomic mosaic of both phylogroups, and that adaptive differences between the breeds would arise from a differential composition of phylogroups at their foundation.

Once the study was completed, I sailed out to optimize two key aspects of the SNP discovery phase: the conversion rate, and the number of SNP identified in the transcriptome. Thus, the second section of this Thesis describes EXFI, a method that uses state-of-the-art algorithms to split transcripts into exons. Transcriptome-derived SNPs achieve low validation rates when the positions of Intron-Exon Boundaries (IEB) are not taken into account. With this problem in mind, Conklin et al., 2013 developed the IEB method, based on the mapping of WGS reads on the transcriptome. Under this approach, WGS reads are mapped to the transcriptome, and places where sudden mapping starts and ends reveal the presence of IEBs.

With this idea in mind, I studied what could happen if instead of mapping reads, I tried to do the same procedure with k-mers. Moreover, I went to state-of-the-art probabilistic data structure to accelerate the procedure even more and also decrease the disk and memory footprint of the method.

The data structure chosen for this task are the Bloom filters. Their first advantage is their speed: checking that an element (k-mers) is in the data structure is fast because it uses hash functions to encode elements as numbers. The second one is memory efficiency: by representing k-mers into digits, a drastic reduction in space can be achieved. Nonetheless, this data structure is probabilistic, meaning that it can produce errors. Concretely, the Bloom filter only makes false positives, with a certain rate, that nonetheless can be kept under control. Moreover, downstream analysis can detect and correct such errors. Therefore Bloom filters are fast, memory-efficient and probabilistic, suitable for the IEB detection task.

The development of this procedure is crystallized in the article "EXFI: Exon and splice graph prediction without a reference genome" (Langa et al., 2020). Within it, I show the algorithmic development of the method, published as a Python3 package. It is a two-step procedure to split a transcriptome into its constituent exons through WGS, rather than a genome assembly.

In the first stage, since transcribable DNA is a small fraction of the genome, the WGS experiment is split between reads that overlap the transcriptome and reads who don't with an auxiliary Bloom filter of the transcriptome. The latter are discarded, when the former are stored into a Bloom filter, made up of k-mers of the WGS reads that overlap the transcriptome.

In the second stage, the transcriptome is split into exons. To do so, each k-mer of the transcriptome is checked in the previous Bloom filter. Each time that a transcriptomic k-mer is missing in the genomic Bloom filter, it indicates that such k-mer is composed of two exons, and therefore is missing in the WGS dataset. With this idea in mind, the transcriptome is splitted each time that k-mers are missing. The structure of the transcriptome is stored as a mathematical graph, where exons represent nodes, the connections between nodes as edges, and transcripts as paths.

To validate the program I used three tiers of datasets. The first one were the human and zebrafish references, because of the amount of knowledge we have in this species, and to guide the initial steps and correctness of EXFI. The second one were fishes, Atlantic salmon and Atlantic herring, because this Thesis studies fishes. Moreover, since the final purpose of this method is to discover SNPs as shown in the first part of this thesis, these datasets were used because there are available Pool-Seq WGS experiments. The advantage of this experimental approach is that by mixing multiple individuals in the same run, it is possible to discover many more variants than doing it individually. The disadvantage is that the number of variants in the dataset might mislead the method. The final tier of species were mega-genomes, the ones from the axolotl and the sugar pine, both almost 30 Gbp in size. The reasoning behind them was to prove that memory efficient algorithms can deal with datasets so large.

Additionally, I compared the performance of EXFI against two tools: ChopStitch, a reference-free method similar to EXFI; and GMAP, a splice-aware aligner, which needs a reference genome in order to discover exons.

There are four aspects that we characterized in EXFI. The first is that the initial WGS read filtering is vital: it reduces by at least an order of magnitude the number of reads that end in the Bloom filter. Therefore, the memory footprint of the Bloom filters can be reduced significantly without sacrificing its accuracy and the ones of the posterior steps.

Second, we searched for an appropriate memory footprint. The error reduction because of the pre-filtering step allows us to decrease the size of the Bloom filter without a significant decrease in accuracy. Therefore, we tested EXFI with a range of Bloom filter sizes from 4 to 60 GB, and achieved almost the same results with the smaller one. In conclusion, 4 GB are enough to process a species in the same order of magnitude as humans and zebrafishes. Therefore, these analyses can be carried out in desktops and laptops.

Third, we determined the optimal k-mer length. A small value of k would reduce the specificity while increasing the number of k-mers to process. On the other side, a large value of k would increase specificity at the cost of a smaller exon discovery rate. Therefore an equilibrium had to be found, which turned out to be between 23 and 35 bp.

Fourth, an appropriate WGS depth needs to be used. If such depth is too low, in some regions could be zero, and therefore would show up as an IEB signal, and EXFI would underperform. On the opposite side, if it is too high, the number of sequencing errors fills unnecessarily the Bloom filter, and therefore the false positive rate increases, and the accuracy drops. To discover the optimal sequencing depth, we sampled the Zebrafish datasets, and discovered that such optimal depth lies in the region of 25 to 30x.

Finally, we compared the performance of EXFI against the two other tools. Generally, with the exception of the *de novo* herring transcriptome assembly, EXFI outperformed the other methods in terms of memory footprint and accuracy of the exon predictions.

Thus, we have shown EXFI, a tool for rapid and effective exon decomposition of a transcriptome, without using a reference genome. A retrospective *in silico* experiment with tench shows that EXFI could have been capable of discovering 228,000 safe-to-genotype SNPs, with a near 100% precision and recall.

The third section of the Thesis shows an optimized sampling method for transcript profiling, with the purpose of increasing the number of transcripts and therefore the number of discoverable SNPs; and at the same time, obtains the most comprehensive transcriptome for the European sardine (*Sardina pilchardus*).

Our work prior in tench has shown us that gene expression is very dependent on the tissues used. If our goal in mind is to discover the maximum number of transcriptome-derived SNPs, we therefore need to maximize the number of expressed transcripts. Therefore, our first goal before sequencing the sardine transcriptome was to discover what tissues contain the most information, and what strategy we should follow to maximize to obtain the maximum number of transcripts.

To do so, we looked at an exhaustive RNA-Seq dataset of twelve tissues of zebrafish (Pasquier et al., 2016). Since this fish is a reference, we could quantify its transcrip expression and also measure the effects of *de novo* assembly. We sampled this twelve tissue dataset under different sequencing depths, and composed *in silico* a mix of all tissues.

Results show that the mix outperformed each individual tissue, even at the risk of fragmenting and losing transcripts due to low coverage.

I applied this multi-tissue strategy to sample, sequence and assemble the most comprehensive transcriptome of the European sardine. The tissues chosen for this effort

were brain, eye, heart, kidney, liver, muscle, ovary, skin, and testes. This sequencing effort was done using the minimum number of donors to minimize the individual sequence variability, which could misguide the *de novo* transcriptome assembly step.

This transcriptome was annotated and quantified.

This transcriptome was vital for the fourth section of the Thesis. Additionally, this transcriptome will be used in the future, using the same approach as in tench, and using EXFI, to discover the population structure of this fish, so important for the ecology and the fishing industry in the Bay of Biscay.

In the last section, I studied the evolutionary relationships within Clupeiformes. They are of great importance to the fishing industry because of their ecological and nutritional value, and lipid and protein content. Nonetheless, these species receive little attention, and until recently, no genome assemblies were available.

The purpose of this last part of this Thesis, under review and titled "Recurrent positive selection of lipid trafficking genes in Clupeiformes" (Langa et al., n.d.), sheds some light over these species. In particular, its purpose was to discover genes under positive selection, trying to discover what genes and biological processes characterize Clupeiformes.

To do so, I collected a large dataset of transcriptomes from twelve species of herrings, anchovies, sardines, and shads; built their phylogenetic tree, and discovered groups of genes under positive selection.

The computational procedure to do so is tremendously complicated, involving multiple clusterings steps, solving paralogy and orthology relationships, processing nucleotide and protein alignments, resolving phylogenies, and producing thousands of evolutionary hypotheses; all to discover genes and biological processes under selection.

We observed under positive selection almost a thousand genes. To discover some structure over them, we looked at the categories of genes, discovering that the groups of genes significantly affected by evolution are 1) the mitochondrial electron transport chain, 2) ribosomes, 3) lysosomes, 4) caveolae, 5) CD molecules, and 6) extracellular proteins.

The main conclusion of this study is that evolution has shaped the Clupeiformes molecular machinery towards an improved storage and transportation of lipids.

In conclusion, this Thesis shows through two case studies that RNA-Seq is a powerful representation of the functional part of the genome, that is cost-effective, and valid for micro- and macro-evolutionary studies.

In the first part, we successfully disentangle the genetic differences between two breeds of tench in Central Europe. Then, we successfully improved the methodology on two fronts. In the first one, we developed a method, EXFI, for an accurate and

efficient exon decomposition, which in turn improves the number of SNPs discovered and conversion rates. In the second one, we designed a transcriptome sampling strategy to maximize the number of expressed and assemblable transcripts, which in turn increases the number of SNPs discovered.

Finally, in the second part, we analyze and describe the evolutionary differences between clupeid fishes, concluding that the most important aspect of their biological machinery is that they have shifted towards an improved transport and storage of lipids.

Resumen

Los recursos genómicos y las herramientas bioinformáticas son muy escasas en especies no modelo tales como la mayoría de peces, y eso a pesar de que haya suficientes razones ecológicas y económicas para que la situación cambie. El RNA-Seq puede ser una herramienta efectiva para generar tales recursos genómicos con el objetivo de revelar la variación genética y funciones necesarias tanto para estudios micro- y macro-evolutivos en especies no modelo. Mi Tesis es la continuación natural de los trabajos previos realizados en mi grupo de investigación. Esta línea de investigacion comenzó con un estudio de genética de poblaciones de la anchoa europea (*Engraulis encrasicolus, L.)*) por Montes et al. (2013), el cual necesitó del desarrollo del método IEB por Conklin et al. (2013). Estos estudios fueron seguidos por una aplicación similar de la metodología en caballa (*Scomber scombrus, L.*; Genomic Resources Development Consortium et al., 2015).

Esta Tesis caracteriza varias especies no modelo de peces en dos diferentes escalas temporales: a nivel de poblacion en *Tinca tinca*, un pez de agua dulce nativo de ríos a lo largo de Europa y Asia Central, y a nivel inter-especies en Clupeiformes, presentes a lo largo del mundo tanto en agua dulce como salada. Ambos estudios usan como cimiento la secuenciación de ARN. Esta Tesis también contribuye con métodos bioinformáticos y *pipelines* para el ensamblaje *de novo* de transcriptomas, y la identificación de (1) sus secuencias, (2) sus funciones, (3) estructura exónica, (4) variación nucleotídica, y (5) genes bajo selección positiva.

La primera sección de esta Tesis, publicada en "A novel transcriptome-derived SNPs array for tench (Tinca tinca L.)" (Kumar et al., 2019), aborda la presencia o ausencia de los filogrupos Este y Oeste de dos poblaciones cultivadas de *Tinca tinca* en Europa Central. Con este fin, la variación genética fue determinada dentro y entre dos razas de la República Checa y Hungría. Este es el primer estudio que genera recursos genómicos y transcriptómicos para esta especie.

Muestreamos RNA de individuos de ambos sexos, dos tejidos (cerebro y músculo) y dos tasas metabólicas distintas (rápida en verano, y lenta en invierno). Ensamblamos un transcriptoma compuesto de 267,058 secuencias, annotándolo con información funcional, y prediciendo el conjunto de secuencias codificantes y tranducción a proteínas. A través de una secuenciación suplementaria de ADN de otros diez individuos, descubrimos 60,414 polimorfismos de nucleótido único (en inglés SNP). De ellos construímos un chip de 96 SNPs de esta especie, el primero, y lo

usamos para genotipar 140 muestras de ascendencia tanto de filogrupo puro como híbrido (este, oeste o híbrida, de acuerdo al gen de la hormona del crecimiento), de dos razas locales de la República Checa y Hungría.

Los resultados de *clustering* indican que el número más probable de ascendencias es dos, el mísmo número de filogrupos, pero la mayoría de individuos tienen una ascendencia mixta. De acuerdo al estadístico F_{ST}, existen diferencias pequeñas pero significativas entre razas, pero no a nivel de filogrupo. Por tanto, dentro de cada raza hay flujo genético entre individuos de ambos filogrupos, y por tanto no hay aislamiento reproductivo entre los dos filogrupos. Nuestro estudio respalda la hipótesis de que los individuos analizados resultan ser un mosaico genómico de ambos filogrupos, y que las diferencias adaptativas entre las razas se deben a la composición inicial de filogrupos en el momento de su fundación.

Una vez que este estudio fue completado, salí a optimizar dos aspectos clave en la fase de descubrimiento de SNPs: la tasa de conversión, y el número de SNPs identificados sobre el transcriptoma. De este modo, la segunda sección de esta Tesis describe EXFI, un método que utiliza algoritmos de última generación para dividir tránscritos en exones. Los SNPs derivados a través del transcriptoma tienen bajas tasas de validación cuando los límites entre intrón y exón (IEB en inglés) no se tienen en cuenta. Con este problema en mente, Conklin et al., 2013 desarrolló el método IEB, basado en el mapeo de lecturas WGS al transcriptoma. Bajo este enfoque, las lecturas WGS son mapeadas al transcriptoma, y los lugares en los que suceden repentinamente muchos comienzos y finales de mapeos revelan la presencia de los IEBs.

Con esta idea en mente, estudié qué podría pasar si en vez de mapear lecturas, intentara hacer lo mismo con k-mers. Es más, acudí en busca de lo último en estructuras de datos probabilistas para acelerar el procedimiento todavía más y también reducir la huella en memoria y disco del método.

La estructura de datos elegida para esta tarea son los filtros de Bloom. Su primera ventaja es su veolcidad: comprobar que un elemento (los k-mers) están en la estructura de datos es rápido porque usa funciones *hash* para codificar elementos como números. La segunda es que es eficiente en memoria: a través de representar un k-mer como números, logrando una gran reducción en el espacio necesario. Sin embargo, esta estructura de datos es probabilista, queriendo decir que puede producir errores. Concretamente, el filtro de Bloom sólo produce falsos positivos, con un ratio predeterminado, que sin embargo podemos mantener bajo control. Es más, los análisis posteriores pueden detectar y corregir tales errores. Por tanto, los filtros de Bloom son rápidos, eficientes en memoria, y probabilistas, apropiados para la tarea de detectar IEBs.

El desarrollo de este procedimiento se cristalizó con la publicación de "EXFI: Exon and splice graph prediction without a reference genome" (Langa et al., 2020). En

él muestro el desarrollo algorítmico del método, publicado como un paquete de Python3. Es un procedimiento de dos pasos que divide el transcriptoma en exones a través de lecturas WGS en vez de un ensamblaje de genoma.

En la primera etapa, puesto que el ADN transcribible es una pequeña fracción del genoma, el experimento WGS es dividido entre lecturas que solapan el transcriptoma y lecturas que no a través de un filtro de Bloom del transcriptoma. Las segundas son descartadas, cuando las primeras son guardadas en un filtro de Bloom, hecho de los k-mers de las lecturas que sí se solapan con el transcriptoma.

En la segunda etapa, el transcriptoma es dividido en exones. Para ello, cada vez k-mer del transcriptoma se busca en el filtro de Bloom genómico. Cada vez que un k-mer del transcriptoma está ausente en el filtro de Bloom, lo que indica es que tal k-mer está compuesto de dos exones, y por tanto no está presente en el experimento WGS. Con esta idea en mente, el transcriptoma es dividido cada vez que falta un k-mer. La estructura del transcriptoma se guarda como un grafo matemático, donde los exones son nodos, las conexiones entre nodos como aristas, y los tránscritos como caminos.

Para validar el programa utilicé tres familias de conjuntos de datos. La primera estaba compuesta por las referencias humana y del pez cebra, por la cantidad de conocimiento que tenemos de estas especies, y para guiar los pasos iniciales y la exactitud de EXFI. La segunda eran peces, el salmón atlántico y el arenque atlántico, porque esta Tesis estudia peces. Es más, puesto que el propósito final de este método es descrubrir SNPs como hemos mostrado en la primera parte, estos conjuntos de datos fueron usados porque hay disponibles datos de secuenciación de Pool-Seq. La ventaja de esta aproximación es que mezclando múltiples individuos en un mismo experimento permite descubrir muchas más variantes que haciéndolo de manera individual. La desventaja es que la gran cantidad de variantes puede confundir al método. La familia final de especies son los mega-genomas, de ajolote y el pino de azúcar, ambos de una longitud de casi 30 Gbp. La razón detrás de ellos era para demostrar que un algoritmo tan eficiente puede procesar conjuntos de datos tan grandes.

Más aún, comparé el rendimiento de EXFI con otras dos herramientas: ChopStitch, otro método libre de referencias similar a EXFI; y GMAP un alineador *splice-aware*, el cual necesita un genoma de referencia para descubrir exones.

Hay cuatro aspectos que caracterizamos en EXFI. El primero es que el filtrado inicial de lecturas WGS es vital: reduce en un órden de magnitud el número de lecturas que terminan en el filtro de Bloom. Por tanto, la huella en memoria de los filtros de Bloom puede reducirse significativamente sin sacrificar tanto su precisión como el de los pasos posteriores.

En segundo lugar, buscamos una huella en memoria apropiada. La reducción de la tasa de error debida al paso de prefitrado nos permite reducir el tamaño del filtro

de Bloom sin una pérdida significativa de precisión. Por tanto, probamos EXFI en un rango de tamaños de filtros de Bloom de 4 a 60 GB, y logramos casi los mismos resultados con el más pequeño. En conclusión, 4 GB son suficientes para procesar una especie del mismo orden de magnitud que el ser humano y el pez cebra. Por tanto, estos análsis se pueden llevar a cabo en equipos de escritorio y portátiles.

En tercer lugrar, descubrimos el tamaño óptimo del k-mer. Un valor de k bajo podría reducir la especificidad a la vez que incrementa el número de k-mers a procesar. En el lado contrario, un valor de k alto incrementaría la especificidad con el coste de una tasa de descubrimiento de exones más baja. Por lo tanto, era necesario descubrir un punto de equilibrio, que resultó encontrarse en el rango de 23 a 35 pares de bases.

En cuarto lugar, buscamos una profundidad de secuenciación WGS apropiada. Si tal profundidad es demasiado baja, en algunos lugares resultará ser cero, y por tanto se mostraría como una señal de IEB, y los resultados de EXFI serían peores. Por otro lado, si es demasiado alta, el número de errores de secuenciación llena de forma innecesaria el filtro de Bloom, y por tanto la tasa de falsos positivos crece, y la precisión cae. Para descubrir la profundidad de secuenciación óptima, muestreamos el conjunto de datos de pez cebra, y descubrimos que tal profundidad óptima se encuentra en la region entre 25 y 30x.

Finalmente, comparamos el rendimiento de EXFI contra las otras dos herramientas. En general, con excepción del ensamblaje *de novo* del arenque, EXFI superó a los otros métodos en términos de huella de memoria y precisión.

Por tanto, hemos mostrado que EXFI, una herramienta para una descomposición rápida y efectiva del transcriptoma, sin utilizar un genoma de referencia. Un experimento *in silico* retrospectivo en tenca muestra que EXFI podría haber sido capaz de identificar casi 228,000 SNPs fuera de peligro, con una precisión y exhaustividad cercana al 100%.

La tercera sección muestra un método de muestreo optimizado para la caracterización de tránscritos, con el propósito de incrementar el número de tránscritos y por tanto el número de SNPs descubribles; y a la vez, obtiene el transcriptoma más exhaustivo para la sardina europea (*Sardina pilchardus*).

Nuerstro trabajo previo en tenca nos ha mostrado que la expresión génica es muy dependiente del tejido utilizado. Si nuestro objetivo en mente es descubrir el mayor número de SNPs basados en el transcriptoma, necesitamos por tanto maximizar el número de tránscritos expresados.

Para ello, echamos la vista a un conjunto de datos de RNA-Seq muy exhaustivo de pez cebra (Pasquier et al., 2016). Puesto que este pez es una referencia, pudimos tanto cuantificar la expresión de sus tránscritos como medir los efectos del ensamblaje *de novo*. Muestreamos este conjunto de datos de doce tejidos bajo distintas profundidades, y creamos *in silico* una mezcla de todos los tejidos.

Los resultados muestran que la mezcla superó a todos los tejidos individuales, incluso bajo el riesgo de fragmentar o perder tránscritos por una cobertura muy baja.

Apliqué esta estrategia multi-tejido para muestrear, secuenciar y ensamblar el transcriptoma más exhaustivo de la sardina europea. Los tejidos elegidos para este esfuerzo fueron cerebro, ojo, corazón, riñón, hígado, músculo, ovario, piel, y testículos. Este esfuerzo de secuenciación fue hecho utilizando el menor número de donantes para minimizar la variabilidad individual a nivel de secuencia, la cual pordría confundir el paso de ensamblaje *de novo*.

Este transcriptoma fue anotado y cuantificado. Este transcriptoma fue vital para la cuarta sección de la Tesis. Es más, este transcriptoma será utilizado en el futuro, utilizando la misma aproximación que en tenca, y utilizando EXFI, para descubrir la estructura poblacional de este pez, tan importante para la ecología y la industria pesquera en el golfo de Vizcaya.

En la última sección, estudié las relaciones evolutivas dentro de los Clupeiformes. Son de gran importancia para la industria pesquera por su valor ecológico y nutricional, y su contenido en lípidos y proteínas. Sin embargo, estas especies reciben poca atención, y hasta hace poco, no había disponible ningún ensamblaje de genoma.

El propósito de esta última parte de la Tesis, bajo revisión y titulado "Recurrent positive selection of lipid trafficking genes in Clupeiformes" (Langa et al., n.d.), arroja un poco de luz sobre estas especies. En particular, su propósito era descubrir genes bajo selección positiva, tratando de descubrir qué genes y procesos biológicos caracterizan a los Clupeiformes.

Para ello, recopilé un gran conjunto de transcriptomas de doce especies de arenques, anchoas, sardinas y sábalos; construí su árbol filogenético, y descubrí grupos de genes bajo selección positiva.

El procedimiento computacional para hacerlo es tremendamente complicado, involucrado muchos pasos de *clustering*, resolviendo relaciones de paralogía y ortología, procesando alineamientos de nucleótidos y proteínas, resolviendo filogenias, y produciendo miles de hipótesis evolutivas; todo ello para descubrir genes y procesos biológicos bajo selección.

Observamos bajo selección positiva casi un millar de genes. Para descubrir algún tipo de estructura en ellos, echamos la vista a las categorías de genes a las que perteneces, descubriendo que los grupos de genes significativamente afectados por la evolución son 1) la cadena de transporte de electrones en las mitocondrias, 2) los ribosomas, 3) los lisosomas, 4) caveolas, 5) cúmulos de diferenciación, y 6) proteínas extracelulares.

La principal conclusión de este estudio es que la evolución ha moldeado la maquinaria molecular de los Clupeiformes hacia un almacenamiento y transporte de lípidos mejorado.

En conclusión, esta Tesis muestra a través de dos supuestos prácticos que el RNA-Seq es una potente represencación de la parte funcional del genoma, que es económica, y válida tanto para estudios micro- y macro-evolutivos.

En la primera parte, desenredamos las diferencias genéticas entre dos razas de tenca en Europa central. A continuación, mejoramos con éxito la metodología en dos frentes. En el primero, desarrollamos un método, EXFI, para una predicción precisa y eficiente de exones, el cual mejora el número de SNPs descubiertos y tasa de conversión. En el segundo, diseñamos una estrategia de muestreo de transcriptomas, la cual aumenta también el número de SNPs descubiertos.

Finalmenre, en la segunda parte, analizamos y describimos las diferencias evolutivas entre clupeidos, concluyendo que el aspecto más importante de su maquinaria biológica es que se ha desplazado hacia un transporte y almacenamiento de lípidos mejorada.

Acknowledgements

Quiero agradecer primero a la Universidad del País Vasco UPV/EHU y al Departamento de Genética, Antropología Física y Fisiología Animal. Doy las gracias al Departamento de Educación del Gobierno Vasco por concederme la beca predoctoral PRE_2017_2_0169.

También quiero agradecer a Andone y Darrell, por darme esta oportunidad de hacer un trabajo tan multidisciplinar en un tema en el jamás hubiera pensado que un matemático tuviera algo que decir: evolución de peces. Ha sido duro, ha costado mucho tiempo, pero aquí estamos por fin.

Gracias a Aitor, porque todo esto fue idea tuya, te fuiste, has vuelto hace nada, has rematado el trabajo, ha salido todo bien, ha quedado bonito, y todavía queda mucha ciencia por hacer.

También quiero dar las gracias a los colaboradores en los papers presentados: a Girish Kumar, por introducirnos en el mundo de las piscifactorías, a Martin Huret y Paul Gatti, por conseguirnos esas muestras de sardinas, y a Yuri Rueda por hacernos de guía en el mundo de las lipoproteínas.

Gracias a la gente del grupo *Abereak/Genomic Resources/Applied Genomics and Bioinformatics*: A los mayorcitos: Mikel Iriondo y Carmen Manzano. A la gente de SGIKER: Irati y Fer. A los que estuvieron pero ya no están: Urtzi, Mikel Egaña, Marta Alfaro, Mikel Aguirre, David, Otsanda, Iratxe Montes, Marta Muñoz, Pablo, Candela y Álvaro. A los que estáis aquí: Melanie, Iratxe Zarra, Iratxe Aguado, Igor y Mikel Gutiérrez. Y especialmente a las dos que me habéis aguantado hasta el último minuto de la recta final: Sofía y June.

Me gustaría agradecer también a la gente del laboratorio de Genética de la Facultad de Ciencia y Tecnología. Primero a Ane, Jon Vallejo, Jone, Iraia, Gartze, Jon Larruskain y Julen, por todas esos cafés, comidas, cenas y fiestas. También me gustaría agradecer a esa gente que está pero que no le gusta tanto lo de salir a cenar: María, Endika, Alejandro, Naiara, Martín, Mahta, Larraitz, Bea y Cristina. Y también a los *bosses* Adrián, Ainhoa, Aintzane, Ana Aguirre, Asier, Josean y Ana Zubiaga.

I want also to acknowledge Dr. Nicolas Salamin from the Department of Ecology and Evolution for allowing me to stay in his group over three months, and also Martha Serrano, for helping and making comfortable the stay, and introducing me to the other students and postdocs. I would also want to acknowledge the organizers of the course Computational and Molecular Evolution held in Crete in 2016, especially Emma and Cilia, because of their patience with us the students, inside and outside the course.

También quiero dar las gracias al grupo de matemáticos de Zaragoza, especialmente a Javi Campos y Juan, que desde la *"aventura"* del Erasmus en Pau, la cosa se empezó

a torcer. También a la Tita Alicia, Adela, Paula, Mar, Diego, Ana, Carlota, Santi, Javi, Luciano, Diego, Josevi.

À Anaïs Baudot et Élisabeth Remy, car si vous travaillez dans un département de mathématiques, je peux travailler dans un département de biologie.

También quiero dar las gracias a la gente del máster de bioinformática, que aunque hayamos terminado todos desparramados por todo el planeta, siempre hay alguna conferencia en la que encontrarnos: Rafa, Cristina, Juan, Fran y Carlos.

Finalmente quiero dar las gracias a mi familia. A mis padres, por preguntarme todas las semanas desde que empecé la tesis que cuándo voy a terminar. A mis hermanos Sebastián y Carmen. Y finalmente a mis sobrinos Mario, Rafael, Bruno y Sofía. A los cinco minutos de veros estoy agotado, pero no sabéis lo bien que me lo paso.

Gracias.

Contents

List of Figures

List of Tables

List of Publications

Published articles:

- Kumar, G., Langa, J., Montes, I., Conklin, D., Kocour, M., Kohlmann, K., & Estonba, A. (2019). A novel transcriptome-derived SNPs array for tench (Tinca tinca L.). *PLOS ONE*, *14*(3), e0213992. https://doi.org/10.1371/journal.pone.0213992
- Langa, J., Estonba, A., & Conklin, D. (2020). EXFI: Exon and splice graph prediction without a reference genome. *Ecology and Evolution*, *10*(16), 8880–8893. https://doi.org/https://doi.org/10.1002/ece3.6587
- Langa, J., Huret, M., Montes, I., Conklin, D., & Estonba, A. (2021). Transcriptomic dataset for Sardina pilchardus: Assembly, annotation, and expression of nine tissues. *Data in Brief*, 107583. https://doi.org/10.1016/j.dib.2021.107583

Articles being reviewed:

- Langa, J., Rueda, Y., Albaina, A., Huret, M., Conklin, D., & Estonba, A. (n.d.). Recurrent positive selection of lipid trafficking genes in Clupeiformes, Manuscript under review on Marine Biotechnology

List of Abbreviations

BF	**B**loom **F**ilter
BF FPR	**B**loom **F**ilter's **F**alse **P**ositive **R**ate
CDS	**C**oding **D**NA **S**equence
DNA	**D**eoxyribo**N**ucleic **A**cid
IEB	**I**ntron-**E**xon **B**oundary
LC-PUFA	**Long**-**C**hain **P**oly**U**nsaturated **F**atty **A**cid
MCMC	**M**arkov **C**hain **M**onte **C**arlo
ML	**M**aximum **L**ikelihood
MYA	**M**illion **Y**ears **A**go
RIN	**R**NA **I**ntegrity **N**umber
RNA	**R**ibo**N**ucleic **A**cid
RNA-Seq	**R**ibo**N**ucleic **A**cid **Seq**uencing
SCO	**S**ingle-**C**opy **O**rtholog
SNP	**S**ingle **N**ucleotide **P**olymorphism
WGD	**W**hole **G**enome **D**uplication
WGS	**W**hole-**G**enome **S**equencing

Dedicado a mis sobrinos Mario, Rafael, Bruno y Sofía.

Part I

Synthesis

Chapter 1

Introduction

The announcement and completion of the Human Genome Project in 2003 was a milestone in Molecular Biology. Its success made interesting and reasonable the assembly of the genomes of many other model species in Genetics: *Drosophila melanogaster* (Adams et al., 2000), *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000), *Mus musculus* (Chinwalla et al., 2002), or *Danio rerio* (Howe et al., 2013).

After years of iterative refinements over the Sanger method (Sanger et al., 1977), a breakthrough happened: parallel High-Throughput Sequencing. This successful methodology cheaply sequences thousands to millions of fragments, at the expense of being shorter and more error prone. At the beginning of this revolution, two commercial sequencers were available: 454 (Margulies et al., 2005) and Illumina (Bentley et al., 2008). Nowadays, two other competitors are delivering sequences of kilobases to megabases with a higher error rate: Oxford Nanopore Technology (Clarke et al., 2009) and Pacific Biosciences (Eid et al., 2009).

The development of these sequencing platforms opened the gates to specific regions of the genome. We have entire new areas and sequencing approaches depending on the focus of research: Metagenomics (Handelsman et al., 1998), Exome-Seq (Hodges et al., 2007), ChIP-Seq (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007), DNAse-Seq (Boyle et al., 2008), BS-Seq (Meissner et al., 2008), RNA-Seq (Wang et al., 2009), Ribo-Seq (Ingolia et al., 2009), Hi-C (Lieberman-Aiden et al., 2009), Single-Cell Sequencing (Islam et al., 2011), and ATAC-Seq (Buenrostro et al., 2013). The development of these techniques has enabled the research community to go beyond describing common patterns of variation in the human genome, as seen in the HapMap (Altshuler et al., 2005) and the 1000 Genomes projects (The 1000 Genomes Project Consortium, 2012) to describe the functional elements within it (ENCODE; Dunham et al., 2012) and the interactions with the microbiome (Turnbaugh et al., 2007).

These breakthroughs would not have been possible without the application and development of bioinformatics, the multidisciplinary field that combines Biology, Computer Science and Mathematics. None of them would have been achievable

without the development of novel algorithms that made possible the processing, integration, comparison and visualization of the massive amounts of data generated.

Although the advances in keystone species (such as primates, research model species, and livestock), the remaining non-model species receive relatively little attention. Given the absence of reference genomes, and the limited resources to study every single species in the Tree of Life, it is necessary to study them effectively using as little as possible.

RNA-Seq covers the expression of the most significant fractions of the genome: those that are transcribed into RNAs, which contain the set of sequences that are translated into proteins (also known as Coding DNA Sequences, CDS). The primary advantage of this approach is that transcribable and coding DNA, the exome, supposes a tiny fraction of the genome, around 1% in Homo sapiens, and therefore we avoid sequencing the rest, which encompasses repetitive DNA, mobile elements and the introns within genes.

Generating a complete genome reference is a colossal economical, personal, experimental and computational effort reserved for great institutions and consortia. Instead, this Thesis focuses on the set of expressed transcripts since it is an information-dense reduced representation of the genome. Therefore, the methods and techniques used in this Thesis rely primarily on RNA-Seq experiments.

However, RNA-Seq has its own complications. First, genes transcribe multiple messenger RNAs, coding or not, and therefore multiple protein products. Second, genes and transcripts express with varying intensities, depending on the cell type, tissue and development stage of the host organism. Therefore, RNA-Seq approaches, whatever the goal in mind is, must deal with these two problems.

Annotation is the characterization of genes and their functions by searching known databases of transcripts and protein sequences. Luckily, a 40% sequence identity is enough to identify with confidence a protein (Rost, 1999) and the gene that produces it. Once identified, we can start assigning functions and its place in the biochemical reactions of the organism.

The most common use of RNA-Seq is to perform differential gene expression. This procedure involves comparing RNA samples from multiple individuals across two or more biological conditions. As different environmental factors stress an organism, cells respond by increasing and decreasing, even silencing, the transcription of RNA. Therefore, RNA-Seq is used to quantify both transcript and gene expression levels. Afterwards, through statistical analysis, we can determine which genes are over- and under-expressed, providing insights into the molecular responses to the stresses introduced.

Yet another use of RNA-Seq data is to discover genomic variation as Single Nucleotide Polymorphisms (SNPs): small mutations, insertions and deletions. As transcriptomes correspond to genes and therefore functional regions in a genome, transcriptome-derived SNPs are informative for adaptive variation (Beaumont & Balding, 2004; Luikart et al., 2003; Morin et al., 2004). Given that most mutations are neutral, and that the coding DNA is under an enormous selective pressure, SNPs in the transcriptome are of great evolutionary importance. Given enough loci, samples and populations, RNA-based SNP discovery is a cost efficient approach to perform population genomics: the micro-evolutionary study of the phylogenetic history, the demography, and the spatio-temporal distribution of genetic diversity within a species.

Finally, given enough genomes and transcriptomes from many species, it is possible to study them also in a macro-evolutionary fashion. Because of the degeneracy of the genetic code, i.e., the redundancy between the 64 codons and the set of the 20 amino-acids and the stop signal, it is possible to discern at the protein level which mutations are silent (they are synonymous -they produce the exact amino acid) from those that produce a change (non-synonymous). By aligning protein-coding sequences, and comparing codon by codon the ratio between nonsynonymous to synonymous mutations, it is possible to analyze the evolutionary pressures that a gene withstands. We can do this for every single gene and therefore observe the effect of evolution over every molecular and biological process. Therefore, RNA-Seq is a powerful tool for the molecular study of the processes and responses within the cell, but also for seeing the effects of micro-evolution of a single species, and the macro-evolutionary forces behind adaptation and speciation of wild organisms.

Thus, this Thesis develops RNA-Seq bioinformatic procedures in two case studies on the characterization of several non-model fish species at two different time scales: at the population level on *Tinca tinca*, a freshwater fish that belongs to the Cypriniformes order, and at the inter-species level in Clupeiformes, the order that contain sardines, anchovies, herrings and shads.

Chapter 2

Aims and Objectives

2.1 Aims

Genomic resources and bioinformatic tools are very scarce in non-model species, and their generation requires the processing of large amounts of genomic and transcriptomic data to build a reference. In this Thesis we had two aims. The first one was to create bioinformatic pipelines for the scientific community that rely mainly on the transcriptome, and that can be applied to any non-model species. The second was to gain insights through state-of-the-art methods for a better understanding of the evolutionary factors driving teleost species evolution in two different time scales: we investigate local and recent adaptations of populations, as well as global and ancient selection events. For these ends, the above mentioned bioinformatic tools and protocols were applied to describe the population structure of the farmable freshwater cyprinid *Tinca tinca*, along with identifying positively selected genes within the economically and ecologically important Clupeiformes order.

Genomic resources are scarce in fish and this situation needs to change. The first reason is ecological: they are present worldwide and occupy the central parts of the trophic web. The second is economical: fish is a rich source of energy and nutrients, and for millenia fishing has been an important human activity. The final reason is statistical: Actinopterygii, the clade of ray-finned-fishes, encompasses 30,000 species, nearly 40% of all vertebrate species. At the beginning of this Thesis only a dozen reference genomes were available at Ensembl, and therefore almost every new study had to start from zero.

Nonetheless, there are economical and material limitations when generating genomic resources, and therefore strategizing is a must. The principal area from which to prioritize resources is the experimental one: we don't need to study the entire genome, but we can make use of reduced representations of it. Exome sequencing is an interesting reduction strategy since it is restricted to exons from protein coding sequences, regions with functional importance. Thus, taken together that most mutations are neutral, and that the protein coding regions of the genome are under very

selective pressure, the observed SNPs are of great importance. The main drawback is that Exome-Seq requires first a known genome to derive capture probes.

Alternative and successful approaches in non-model species are Genotyping by Sequencing (GBS; Elshire et al., 2011) and Restriction Associated DNA marker Sequencing (RAD-Seq; Baird et al., 2008). Both approaches rely on restriction enzymes and fragment size selection in order to sequence a small fraction of the genome, and then extract polymorphic markers. Their drawback is that the regions studied change from species to species, making impossible comparative studies, and that most of them will lie in intergenic and intronic space, and therefore their associated function, if any, can be hardly attained.

The works presented in this thesis, along with the ones they are based on (Conklin et al., 2013; Genomic Resources Development Consortium et al., 2015; Montes et al., 2013), combine the use of reduced representation of the genome, the transcriptome, and shallow Whole Genome Sequencing (WGS), to obtain the same advantages of exome sequencing without a reference genome, and the same aims: to study the genetic variation exclusively in functional regions, unlike GBS and RAD-Seq.

In conclusion, transcriptome-derived SNPs are more informative and meaningful compared to the reduced representation alternatives because they provide genetic variation and functionality at the same time. Since this is not a standard approach, there are no bioinformatic tools tailored for it, and therefore we first invested time developing them.

The hypothesis of this Thesis can be resumed as the following:

> RNA-Seq is an effective tool for generating genomic resources aimed at revealing genetic variation and function necessary for both micro- and macro-evolutionary studies in non-model species.

Which can be divided in:

1. RNA-Seq is a valid tool to study the local evolution of two cultured breeds of *Tinca tinca*.
2. Transcriptomes derived from RNA-Seq can be split into its constituent exons.
3. Gene representation in the transcriptome can be maximized applying an appropriate sampling strategy.
4. RNA-Seq is a valid tool to identify genes under positive selection
5. No reference genome is required.

2.2 Objectives

To reach the aims, we divided this Thesis into five objectives:

Objective 0: Pipelines

In this objective we built the computational background in which the Thesis is supported. It consisted in:

a) To test and develop computational rules and protocols to construct pipelines that are reproducible, automated, scalable, and easy to maintain and execute.

b) To construct the pipelines relevant for this thesis: transcriptome assembly and annotation (common to all objectives), SNP discovery (Objectives 1 and 3, validation of EXFI (Objective 2); and gene clustering, species tree construction, and search for positive selection (Objective 4).

Given the support roles of this objective, no explicit publication was achievable, but its impact has made possible a dozen of repositories of automated pipelines, all four publications resulting from this THESIS, and published works outside it.

Objective 1: Tench

The purpose of this objective was to design an array of SNP markers for population genetics studies in the *T. tinca* species, and implement it to unravel the phylogroup composition of tench cultures in Central Europe. To achieve this objective the following milestones were considered:

a) To assemble *de novo* the *T. tinca* transcriptome.

b) To perform variant discovery by mapping WGS reads to the transcriptome, and to identify suitable SNPs for genotyping, avoiding Intron-Exon Boundaries.

c) To genotype a subset of the discovered SNPs over a larger number of samples, and apply statistical approaches to reveal the genetic structure of *T. tinca* in Central Europe.

The paper entitled "A novel transcriptome-derived SNPs array for tench (Tinca tinca L.)", (Kumar et al., 2019), co-authored with Dr. Girish Kumar, from the University of South Bohemia in České Budějovice, described and explained the genetic diversity observed within and between the Czech Republic and Hungarian cultivated breeds.

After the study of population genetics in tench, and since the sequencing depth increased both for the RNA-Seq and WGS experiments compared to previous publications, I wanted to improve the results of future experiments. On the one hand, I wanted a more scalable method capable of dealing with more transcripts, depth and samples. On the other hand, I wanted to achieve an almost complete transcriptome with the smallest sequencing effort. This motivated Objectives 2 and 3, respectively.

Objective 2: EXFI

The increasing throughput of the sequencing machines motivated us to develop a new method for transcriptome decomposition into exons, which in particular provides a solution to the IEB (Intron-Exon Boundary) problem. The tasks performed were:

a) To develop EXFI, a tool to decompose transcriptomes into exons without a reference genome, and assess the effects of genome size, coverage, heterozygosity, memory consumption, speed, and accuracy.

b) To modify the SNP discovery procedure from Objective 1b using EXFI, comparing the success of the newly discovered SNPs with the ones used in *Tinca tinca*.

The publication of "EXFI: Exon and splice graph prediction without a reference genome" (Langa et al., 2020), reported the success of the EXFI approach on a wide number of datasets, and its advantages over other tools on most of the important metrics. In addition, it showed the enhancement of SNP analysis in *T. tinca*.

Objective 3: Sardine

Given the success of SNP discovery in *T. tinca*, we also desired to obtain a cost-effective strategy for RNA-Sequencing. To do so, we wanted:

a) To determine a cost-effective sequencing strategy to obtain an almost complete transcriptome by means of simulating RNA-Seq experiments in zebrafish, varying both the sequencing depth and tissues used.

b) To apply the developed strategy to construct the *Sardina pilchardus* transcriptome.

The results of this objective were published in the article "Transcriptomic dataset for Sardina pilchardus" (Langa et al., 2021).

Objective 4: Clupeids

In this objective we applied a phylogenetic procedure to 12 species of Clupeiformes to identify positive selection in these species. The tasks to be done were:

a) To explore public datasets of transcriptomic experiments, and assemble and annotate the twelve clupeid transcriptome sequences available, including the one from *S. pilchardus* above.

b) To cluster the species' transcripts into orthogroups, and use that information to compute the species tree in an attempt to elucidate the phylogenetic relationships within Clupeiformes.

c) To discover genes under positive selection that characterize Clupeiformes, and from these, unravel which functions have been evolutionarily altered in terms of Gene Ontologies, pathways, and gene families.

A list of positively selected genes was provided in the Clupeiformes family, and molecular mechanisms under selective pressure will be reported in the manuscript *Recurrent positive selection of lipid trafficking genes in Clupeiformes* (Langa et al., n.d.).

Chapter 3

Workflow and Theoretical Framework

This Thesis is divided into two major evolutionary themes. The first is the population genetics study of tench (Objective 1), followed by two methodological improvements: optimized exon prediction (Objective 2) and better gene sampling (Objective 3). The second theme is the macro-evolutionary study of Clupeiformes (Objective 4). The outline of this Thesis is illustrated in Figure 3.1.

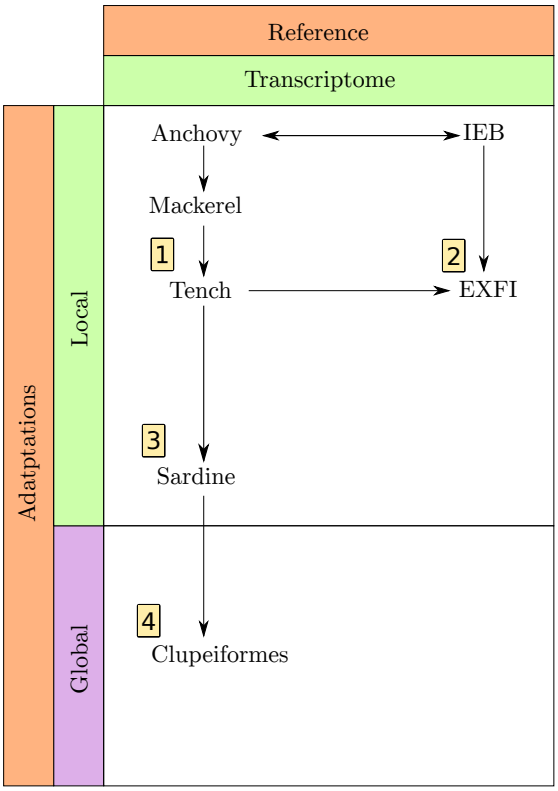


FIGURE 3.1: Outline of this Thesis. Figure 5.7 in Chapter 5 revisits this diagram with applications and parallel publications.

My Thesis is a natural continuation of previous works carried out in my research group. It starts from the population genetics study of the European anchovy (*Engraulis encrasicolus, L.*) by Montes et al. (2013), which needed the development of the IEB method by Conklin et al. (2013). These studies were followed by a similar application of the methodology on Atlantic mackerel (*Scomber scombrus, L.*; Genomic Resources Development Consortium et al., 2015) and tench (Objective 1). After studying the latter fish, two methodological novelties were implemented: better exon prediction with EXFI (Objective 2) and wider gene sampling on Atlantic sardine (Objective 3). Then, evolutionary relationships on Clupeiformes were studied (Objective 4).

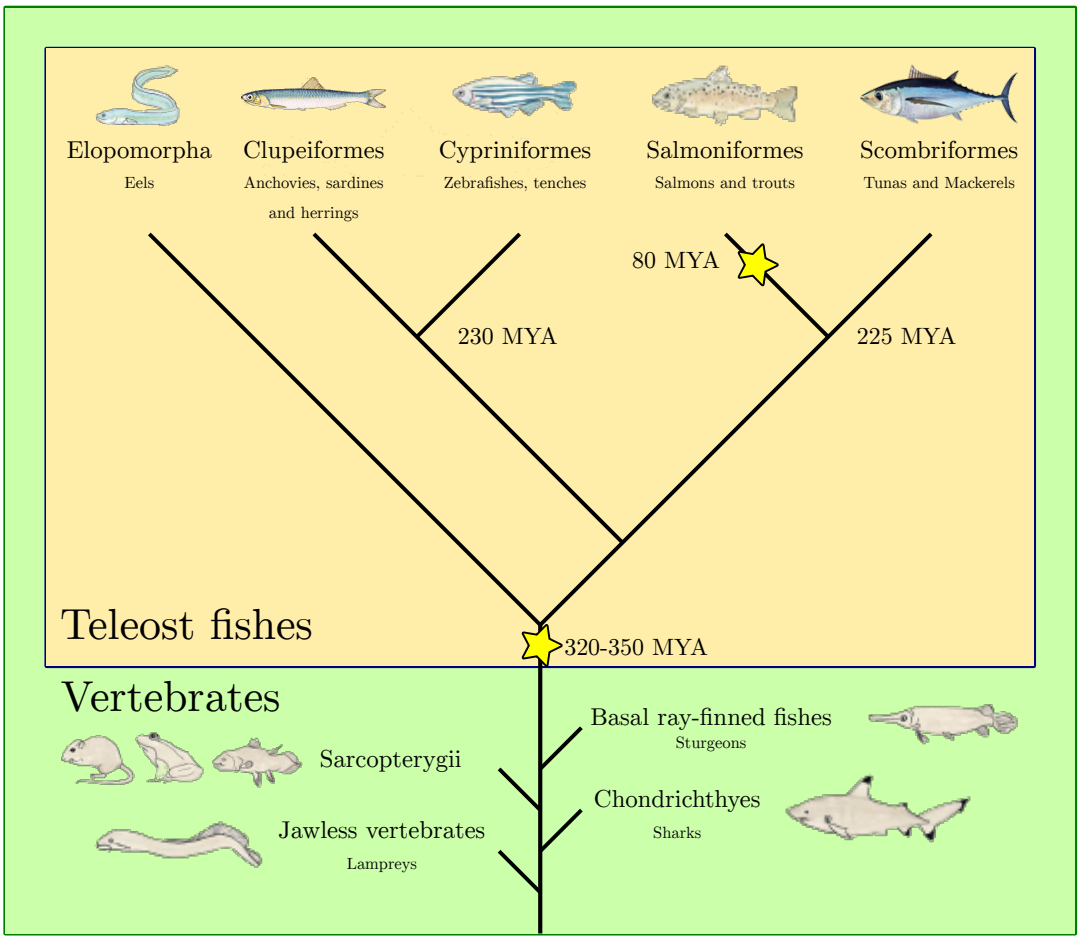


FIGURE 3.2: Simplificated phylogeny of vertebrates and teleost fishes. Denoted with a star are the Whole Genome Duplication rounds occured in some of these species: the Teleost Specific WGD, dated 320 to 350 MYA, and the Salmonid-Specific WGD, 80 MYA. Adapted from Glasauer and Neuhauss (2014).

Teleost fishes are the infraclass that comprises 96% of the fish species in the world, almost half of all vertebrates, and encompasses about 40 orders and 448 families. Among many particularities, the teleost genomes have undergone a Teleost-Specific Whole Genome Duplication (WGD), occurring 320 to 350 million years ago (mya;

Christoffels et al., 2004; Vandepoele et al., 2004). We can find additional WGD events in the teleost tree, for example, one common to Salmoniformes (50 to 80 mya; Figure 3.2; Alexandrou et al., 2013). Although these duplications double the gene content of fish genomes, over time it has resulted in the loss of most genes, with neo- and sub-functionalization of the surviving new copies. This means that fish are prone to have duplicated genes regarding vertebrates, in particular human and mouse, the two species from which the scientific community derives all molecular information. Zebrafish (*Danio rerio*) are one of the central model species in developmental biology because of the robustness of its embryos and their transparency, and therefore exist exhaustive molecular resources for this species. For these reasons, zebrafish are the fish reference species by excellence, and so the reference point in this Thesis.

3.1 Software Development

The evolutionary studies carried out in this Thesis require numerous bioinformatic tools to deal with transcriptomes of non-model organisms. These studies also involve an elevated number of species. Therefore, it was necessary to develop a framework to automate all computational tasks. To do so, I organized the task in different Snakemake workflows (Köster & Rahmann, 2012), rather than a collection of scripts. The reasoning behind was to reduce developing and maintenance costs, ensure the reproducibility, improve the scalability of the constituent steps (both in terms of samples processed and computing resources), and to allow the recycling and their applicability to experiments for any member of the research community. In a commitment to reproducibility and open science, I have published all workflows and datasets in their corresponding paper and form part of public repositories.

3.2 Transcriptomics and Population Genetics: Application to *Tinca tinca*

Tench is a freshwater fish species within the Cyprinidae family, the same that contains zebrafish. It is native to Eurasia, but because of human-mediated movement, tench inhabits temperate and tropical freshwater regions (Welcomme, 1988). Its appearance and flavour makes tench a common fish in aquaculture and sport fishing (Kocour et al., 2010). As of 2019, the annual global aquaculture production of tench is about 1,400 tons (FAO, 2019).

Previous studies have revealed the existence of two phylogroups in Eurasia: a Western one present from the British Isles to Poland, and an Eastern one that goes from Central Europe to Central Siberia and China (Kocour & Kohlmann, 2011; Lajbner et al., 2011). Both phylogroups overlap in Central Europe, where individuals have undergone natural and human-aided hybridization, and these hybrids now appear both in natural and cultured stocks across Europe.

The **first objective** of this Thesis was to determine the genetic structure of tench in two Central European fish farms, focusing mainly on the known phylogroups of tench origin, based on the sequence variability of the transcriptome.

Conklin et al. (2013) provide the basis for a successful method already applied to other fish species based on transcriptomes (Genomic Resources Development Consortium et al., 2015; Montes et al., 2013). A SNP array was designed and applied to disentangle the population structure of two cultured tench breeds native to the Czech Republic and Hungary (Kocour & Kohlmann, 2014).

Once we reached this goal, we retrospectively analyzed two limiting factors of the methodological approach used in the study. First, the bioinformatic approach developed to safely genotype transcriptome-derived SNPs, adapting it to the increasing throughput of sequencing machines year by year. Second, the sample choice needed to be optimized since both coverage and tissue sampling affects the number of genes discovered, and therefore the number of identifiable SNPs.

3.3 Towards an Exome Decomposition without a Reference

To tackle the first problem, we extended the approach followed in Conklin et al. (2013). In a nutshell, it discovers IEBs by mapping genomic reads to the transcriptome. Positions where multiple alignments suddenly start and end are potential candidates for an IEB, because genomic reads are part exon, part intron (Figure 3.3). Following this approach, transcripts are split into exons, and the mapped reads are used to call SNPs.

We can integrate multiple improvements to the procedure. First, we can achieve similar results by mapping the constituent k-mers of the read, simplifying the alignment step. Second, mapping k-mers is much faster than mapping entire reads. Third, since the transcriptome is a small fraction of the genome, a fast filtering criterion can speed up the procedure prior to mapping. And fourth, with the correct data structure, we can apply the inverse procedure: to search for transcriptomic k-mers in the WGS experiment and look for the gaps as hints of IEBs. We looked to probabilistic data structures (also known as sketches), a set of tools that have been applied to bioinformatic problems in the last decade, concretely the Bloom filter (BF, B. H. Bloom, 1970).

The purpose of the **second objective** was the development of EXFI (Langa et al., 2020), initials from exon finder, a tool that uses Bloom filters to decompose a transcriptome into its constituent exons in form of splice graph. Under this representation, 1) nodes are exons, 2) transcripts are paths, and 3) connected components are genes.

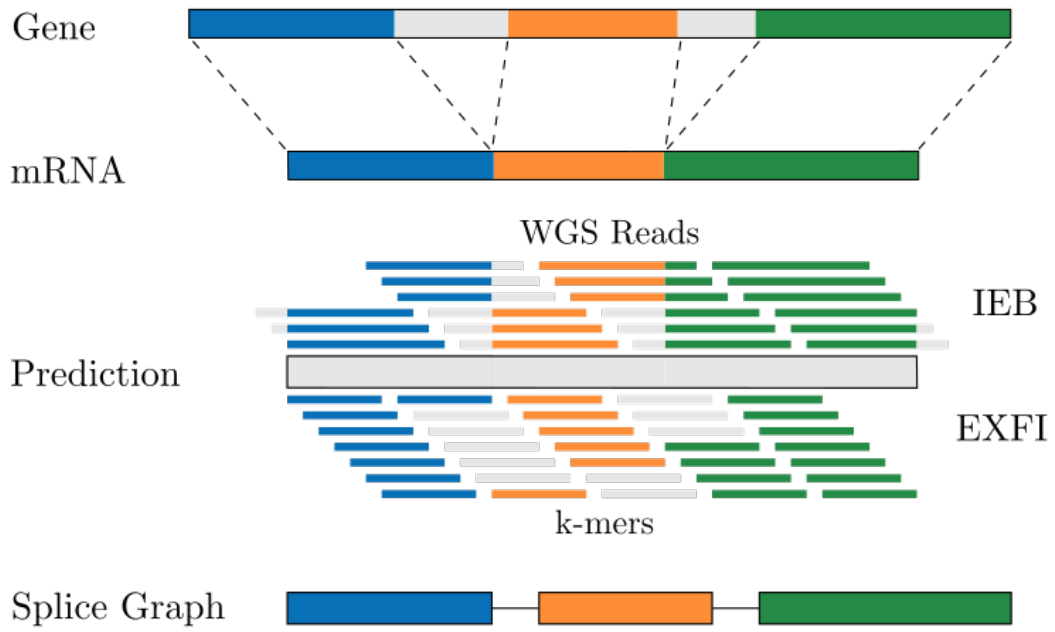


FIGURE 3.3: Summary of the IEB and EXFI procedures. Genes are represented as exons (in different colors) separated by introns (in grey). When genes are transcribed only exons are observed, but their boundaries are unknown. To predict the distribution of exons, either you can 1) *de novo* assemble a genome and use a splice-aware aligner (the standardapproach), 2) map WGS reads against the transcriptome and find out places where alignments are possible (IEB), or 3) check every RNA k-mer in a DNA database (EXFI).

3.4 Towards a Full Transcriptome Assembly: a Simulation in Zebrafishes and an Experiment in Sardines

The second problem we faced on transcriptome-derived SNPs was the RNA library representation. Gene and transcript expression depends on factors such as environmental stresses, the cellular and tissue type, or the developmental stage of the host. Moreover, gene expression follows an exponential distribution: a few genes overwhelmingly express more transcripts than the rest, completely eclipsing the experiment (see Figure 3.4). Additionally, each tissue expresses a fraction of the genes to function, some of them specific to it. Also, we have an economical and experimental constraint: it is not possible to sample every single tissue in depth. On the opposite side, insufficient sequencing results in fragmented and missing transcripts, or choosing a single sample will end in a limited view of the full picture, or including too many individuals will difficult the assembly due to the confusion introduced by the sequence variation of every individual. Therefore, an experimental design has to be found that maximizes gene and transcript discovery and assembly, while constraining sequencing depth, experimental work, and samples used. To discover such a strategy, we proceeded with a two-step approach: a simulation and a real

experiment.

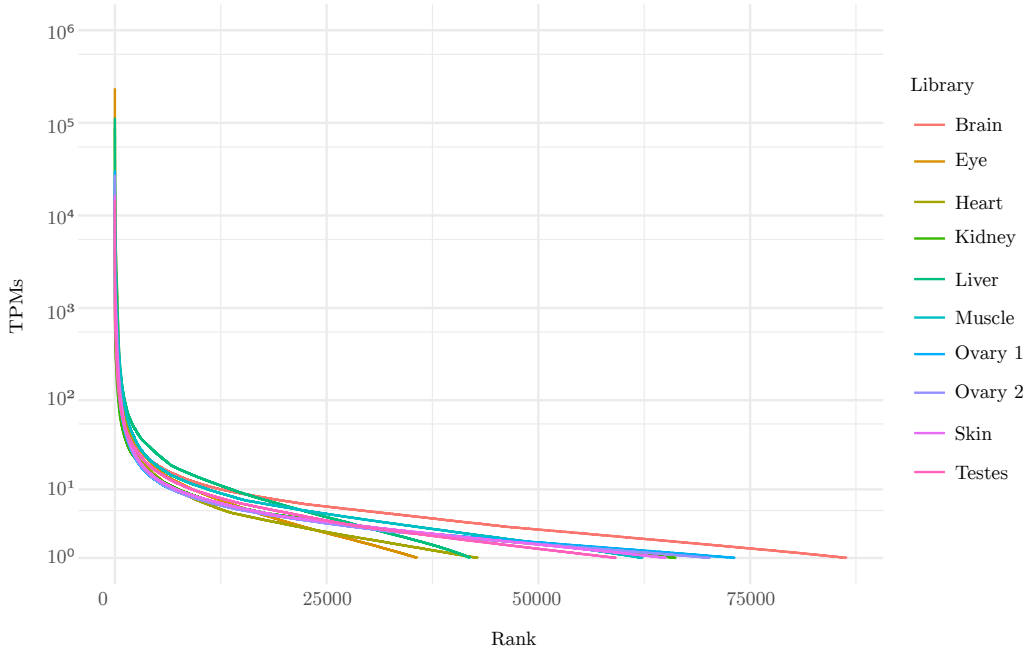


FIGURE 3.4: Ranked normalized expression of transcripts in ten tissues of the European sardine. Units are in Transcripts per Million (TPMs, the ratio per millionth of reads that belong to a gene. Only genes with expression above one TPM are shown. Note the logarithmic scale in the y axis. The eye library (ochre), after normalization, has the highest expressed gene (Gamma-crystallin M2, almost one fourth of reads) and is the least variable of them. On the other hand, the brain (red) has the widest expression profile, and therefore is a good candidate for gene discovery. Nonetheless, the assembly contains almost 200,000 transcripts, and therefore the brain only covers half of the landscape.

For the simulation step, since the purpose of this Thesis is centered on fish, and the keystone fish species is the zebrafish, a dataset from this species was used. Originally published in Pasquier et al. (2016), it is composed of libraries collected from twelve tissues, making it, to my knowledge, one of the most complete and deep in fishes. Since each tissue provides different amounts of information, multiple strategies were prepared in terms of depth and tissue sampling. The resulting optimal strategy was applied to sequence the transcriptome of the European sardine.

The European sardine (*Sardina pilchardus*, Walbaum 1792) is a small pelagic foraging fish that dwells the East North Atlantic coast, from Senegal to the North Sea, and the Mediterranean and Black Sea. It is of great economical and nutritional importance in European countries where climate change and overfishing represent the most significant threats to this natural resource (Checkley et al., 2017). The situation is so dire that Portuguese and Spanish governments imposed a fishing moratoria in 2018 (ICES, 2018). In order to help in the sustainable management of this species

and assess the recovery of the sardine and its populations, it is essential to generate genomic resources.

The aims of the **third objective** of this Thesis were twofold. First, was to discover an optimal experimental strategy for RNA-Seq that maximizes transcript discovery under a limited sequencing budget, and second, to characterize the transcriptome of the European sardine using it. To achieve them, we sequenced and assembled *de novo* nine sardine tissues. We quantified and annotated the transcriptomic sequences and provided information about homology to known proteins, Gene Ontologies, and associated metabolic processes.

Objective 3 was achieved with the publication of Langa et al. (2021), where we published the RNA-Seq reads, transcriptome assembly, annotation and the expression profiles of nine tissues of sardine. The description done will contribute to future research on this species. Moreover, it will be used to characterize the sardine population structure across the European coast, as done in Objective 1. Finally, this dataset was fundamental to the results of Objective 4.

3.5 Evolution of Clupeiformes

Since my research group has been studying for years small pelagic fishes (Albaina et al., 2016; Genomic Resources Development Consortium et al., 2015; Huret et al., 2020; Montes et al., 2013; Zarraonaindia et al., 2012), it was the moment of going beyond population studies and researching what are the genetic features that characterize these species so important to the fishing industry in the Bay of Biscay, in particular clupeids. Therefore, the fourth objective of this Thesis was the comparative study of Clupeiformes.

Clupeiformes is the order of teleost fishes, composed of over 400 species, that includes anchovies, herrings, allis, shads and sardines (D. Bloom and Egan, 2018; Figure 3.5). They are found across the globe in tropical and temperate latitudes, both in salt and freshwater. Ecologically, they are keystone species that mediate between the plankton at the bottom and the predators at the top of the trophic web, either other fishes, marine mammals and seabirds.

According to the Food and Agriculture Organization (FAO), herrings, anchovies, and sardines supposed 24% across all reported fish catches in weight in 2018 http://www.fao.org/fishery/statistics/global-capture-production/query/en. Mainly because of overfishing and global warming, the stocks of these species are decreasing, periodic collapses are reported, and subsequent fishing bans are enforced (ICES, 2018). A better understanding of Clupeiformes biological features is necessary to achieve a sustainable management of these species. Undoubtedly, generating genomic resources for these species will contribute to this understanding.

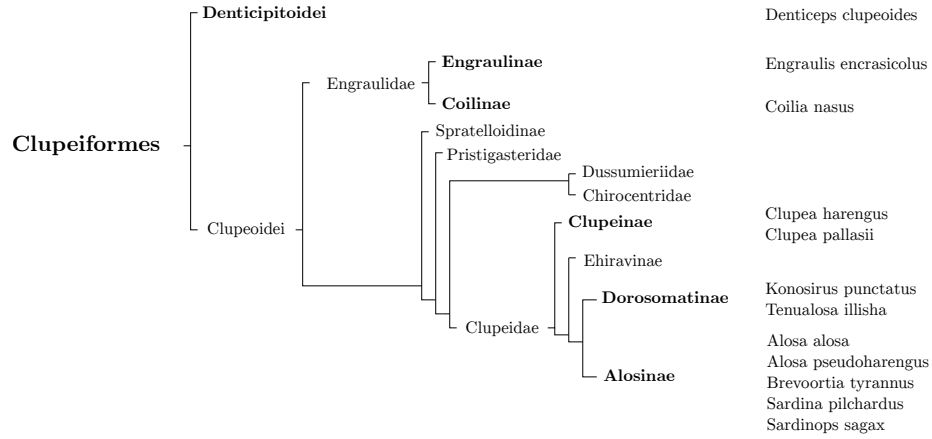


FIGURE 3.5: Simplified phylogeny of Clupeiformes, based on the results of Lavoué et al. (2013)). On boldface are indicated the suborders, families and subfamilies of interest. The right column holds the name of the twelve species studied in this Thesis.

One of the key reasons for the consumption of fish in general, and clupeids in particular, are their high concentrations of protein and long-chain polyunsaturated fatty acids (LC-PUFAs), especially the eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) ω-3 fatty acids, linked to positive health benefits (Burdge, 1998; Calder, 2010; Jain et al., 2015; Kim & Mendis, 2006; Lemaitre et al., 2003; Ruxton et al., 2004; Sidhu, 2003; Uauy & Valenzuela, 2000; Yokoyama et al., 2007). Therefore, a key understanding of the clupeid transcriptome would shed light on how these species synthesize, transport and store fats and these LC-PUFAs.

Little genomic resources existed for these species when this Thesis started. Since then, the situation has improved. In 2016, Barrio et al. (2016) published a draft genome assembly for the Atlantic herring. Then, Machado et al. (2018) and Louro et al. (2019) released two genome drafts for the European sardine. Finally, in the first phase of the Vertebrate Genome Project, Rhie et al. (2021) published a complete assembly for the denticle herring (*D. clupeoides*). As we can see, reference genome assemblies are still reserved for big consortia.

Nonetheless, there are still not enough genomes to characterize these species. Instead, I collected and analyzed publicly available RNA-Seq studies, and used them to integrate them with ours to produce the results for this objective.

While the first objective was centered on recent and local adaptations of a single fish, the **fourth objective** is ancient, planet-wide evolution of Clupeiformes and therefore, the SNP comparisons between populations done for Objective 1 cannot be extrapolated to the larger differences that happen between species. Instead, I looked toward phylogenetic tests, suited for the major variations between homologous genes, and that consider their evolutionary histories. These tests are tailored

for coding sequences, and consider the amino acid changes introduced by mutations at the codon level. We used the branch-site test (Zhang et al., 2005) to detect positive selection in concrete genes and branches of the Clupeiformes phylogenetic tree. The application of this procedure to every single gene and branch allows us to reveal evolutionary patterns with no previous hypothesis in mind, and therefore unbiased towards a goal in particular. As these evolutionary tests work on the codons of protein-coding genes, it is unnecessary to assemble entire genomes to produce them: only transcriptomes.

Chapter 4

Materials and Methods

This section describes the Materials and Methods of the different articles described. It follows the same theme and structure of the objectives: a software development part, followed with the population study of tench, the presentation of EXFI, the transcriptome sequencing effort in sardine, and the comparative study in Clupeiformes.

4.1 Software Development and Protocols Designed

Due to the number of samples and experiments processed, and the computational approach taken, it was necessary to adopt a robust software development approach. Among the many approaches available, we explain the principal ones that ensure automation and reproducibility: Workflow Management and Version Control Systems. Next, the two most used pipelines used in the thesis, *de novo* transcriptome assembly and annotation are explained in detail.

4.1.1 Workflow Management Systems: Snakemake

First, we embraced Workflow Management Systems (WMS). As the number of computational steps and number of samples increases, the approach of copying and pasting commands to the terminal, or writing a great amount of shell scripts, do not work and do not scale up computationally and cognitively. In a nutshell, computational approaches consist of refining massive amounts of raw data into smaller high quality datasets, to then obtain a few tables and figures ready for publication. In definite, a set of files enter, intermediate files are generated, and final outputs (tables and figures) are produced. Therefore a computational pipeline is a set of rules with inputs, outputs and how to convert ones into others. Originally, Make was developed in the 1970's to compile programs and libraries specified in a file usually named Makefile. This file contains rules on how to convert an input into an output. The user only needs to provide the name of the required files and Make takes care on how to schedule all the necessary tasks. Although effective, bioinformatic pipelines exist as Makefiles, but today the main WMS are Snakemake (Köster & Rahmann, 2012), based in Python, and Nextflow (Di Tommaso et al., 2017), based in Groovy.

Both systems are similar, but I decided to focus on Snakemake given my knowledge of Python and its simplicity compared to any other programming language.

Snakemake rules are similar to the ones in Make: in their most simple form, they are defined by specifying the input and output file names and the Shell, Python or R code to convert the one into the other, and it will take care if they have to be generated, in which order they need to be scheduled, and how to execute them in parallel, either in a single server or in a cluster. Figure 4.1 shows a small rule for read mapping.

```
rule map_bwa:
    """Map a sample to reference"""
    input:
        genome = "genome.fa",
        reads = "{sample}.fastq",
    output: "{sample}.bam"
    threads: 32
    conda: "map.yml"
    shell:
        "bwa mem -t {threads} {input.genome} {input.reads} "
        "| samtools sort -o {output} /dev/stdin
```

FIGURE 4.1: Example of a Snakemake rule. This rule maps reads to a reference genome and converts the results into a sorted BAM file. The inputs are the reference genome and a set of reads in FASTQ format. The number of threads used is dynamically chosen between the ones specified and the number available in the computing node. The conda parameter specifies the environment where the needed tools are. The shell parameters contain the explicit chain of commands to execute, substituting the wildcards in curly brackets into the corresponding values specified before.

The first reason to use Snakemake is its integration with the Conda package manager, which in turn grants access to all bioinformatic software available in Bioconda (Grüning et al., 2018). In addition, it integrates seamlessly with major workload managers present in computing clusters such as SGE and Slurm, accelerating the execution and avoiding the user to write and schedule job scripts manually. The final advantage is the containerization of the pipeline, i.e., the virtual isolation of the pipeline with respect to the Operative System, ensuring reproducibility on other machines through Singularity containers (Kurtzer et al., 2017). Moreover, full containerization can be achieved using Docker.

Due to the reasons enumerated before, most of each objective and each paper of this Thesis can be obtained through Snakemake workflows, which are available at GitHub and archived in public repositories (Table 4.1).

4.1.2 Version Control Systems and Git

As discussed with Snakemake, it is impossible to leave the complexity of any bioinformatic project as a monolithic script subject to changes at any given moment, either when writing what steps to execute or, once a mistake is made, what has to change. Therefore, a safer and more effective way to store the methods than overwriting a file, or renaming it with an ever increasing number of suffixes, is not the way to go. To solve this gap in the workflow, I looked at Version Control Systems.

Second, we embraced Git as the Version Control System (VCS). Since the same computational procedures are applied to different samples, it was necessary to have all experimental methods synchronized and updated to use the exact same program versions, databases and parameters.

A second reason to use a VCS is that it keeps the history of all changes made. With this property in mind, it is possible to experiment with the code, add the new contributions when they are successful, and delete them safely when not. Moreover, the history of the project can be divided into branches, with one typically named "master" or "main" used for code ready to be distributed and used, and another one called "devel", where new additions are written, but are not yet ready to be used. This separation allows us to play and experiment with the code safely.

A third reason to use a VCS in general, and Git in particular is the use of hooks, scripts that are automatically executed every time that the developer tries to store changes. Each time that a change is introduced, a battery of tests ensure that the analysis works properly, and therefore the changes are safe to store, and refuses to save them when not. This therefore forces us to check every individual component (unit testing) and to introduce small datasets to prove that the methods work. Through unit testing, we liberate the developer of the cognitive load of which piece does what anytime, and we assure the final user that the methods can be trusted since each one of its components and the overall analysis work.

Apart from unit tests, a complementary approach is to lint the scripts: to unify the writing style, flag correct but misleading code or prone to error. Although simple, this approach removes too cognitive load from the developer, and allows third party inspection of the code since there is only one coding style to use. Additionally, code linters mark duplicated code and sections that seem too complex to understand. Therefore, if code is not simplified, no changes can be stored, and therefore linting promotes the factorization of programs into simple, understandable and maintainable code.

As an example, the software developed for Objective 2, EXFI, is a Python3 package versioned with Git, available at GitHub. The master branch is the one with the ready to use package, the devel branch is the one that contains the set of new features to be merged with master, and from devel multiple branches appear: experiments made to the code to improve it. Some of these experiments were successful and were

TABLE 4.1: Pipelines designed for this thesis.

Name	Purpose
smsk	Boilerplate for other pipelines
smsk_khmer_trinity	Trimming, Normalization and Transcriptome assembly of Ilumina reads
smsk_454	Quality control and transcriptome assembly of 454 reads
smsk_trinotate	Transcriptome annotation
smsk_snptrans	RNA-Seq and WGS based SNP discovery
smsk_exfi_validation	Battery of tests for EXFI, ChopStitch and GMAP
smsk_selection	Transcriptome clustering, tree species construction and detection of positive selection

merged back to devel, others were a failure and were abandoned, but nonetheless they still are present in the history of the project. With respect to continuous testing, multiple unit tests are done for every single function of the package, no matter how big or small they are. Additionally, a small dataset is provided to test that the analysis the package carries out is done from beginning to end. Also, to ensure that the code style is correct and that there are no opportunities for any mistake, it is checked with *pylint*, which checks errors or even the opportunity for errors, forces to use coding standards, and points out code that need to be refactored, i.e., simplified for better understanding.

In conclusion, working with a VCS such as Git is useful not only to effectively save and publish the computational methods, but also to enforce robust good practices. This procedure is crystallized in the smsk (Snakemake skeleton), a template from which all other pipelines are derived to jumpstart them (Table 4.1).

In the following subsections we explain the two most used pipelines across the thesis: transcriptome assembly, and transcriptome annotation.

4.1.3 *smsk_khmer_trinity*: Transcriptome Assembly and Quality Assessment

Since *de novo* transcriptome assembly is vital to this thesis, and that it would be performed across all objectives, an automated pipeline was constructed to convert RNA-Seq reads into a transcriptome without human intervention. This pipeline is composed of three stages: trimming, normalization and *de novo* assembly (Figure 4.2).

Trimming consists of cutting or deleting reads containing adaptors and low quality sequences. The program chosen for this task was Trimmomatic (Bolger et al., 2014) due to its high performance, the facility to personalize the cleaning procedure, and its ability to remove adaptors effectively.

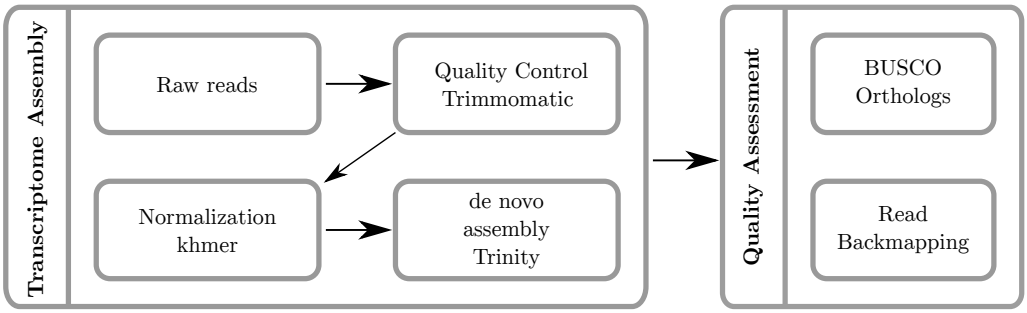


FIGURE 4.2: Schematic representation of the transcriptome assembly pipeline. Reads are cleaned with Trimmomatic, normalized with khmer and assembled with Trinity. Quality assessment is performed by backmapping the reads to the transcriptome and by searching SCOs with Busco.

The cleaning procedure worked as follows: For every read, remove adapters that match the ones used in the TruSeq3-PE-2 protocol. Then, remove bases at the 3' end that have a probability of error below 99% (Q20). Also, remove bases at the 5' end that have a probability of error below 99% too. Then, compute the mean probability of error of the read. If it falls below 99.9%, remove it. Finally, if a read has a length less or equal to 32, remove it. This entire procedure is encoded in a single string as follows: `"ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:20 TRAILING:20 AVGQUAL:30 MINLEN:32"`. Therefore, Trimmomatic has a very powerful syntax to specify what to remove and in which order that other programs lack.

Next, reads were normalized with khmer (Crusoe et al., 2015). This software uses the count-min sketch (Cormode & Muthukrishnan, 2005), a probabilistic data structure used to compute the frequency tables of the k-mers of an experiment. Once computed this frequency table, every sequence can be reprocessed to check the frequencies of its constituent k-mers. If the frequency of a k-mer is too low, that k-mer is considered a sequencing error, and therefore it is removed from the experiment. On the other hand, if a k-mer is too frequent (repetitive regions in WGS, overexpressed transcripts in RNA-Seq, dominant species in Metagenomics), it is redundant to include all the copies, and it is therefore necessary to reduce them, even to delete them all. Therefore, this procedure is very useful to both remove errors and redundant fractions of the experiment at hand, which will result in a faster and more precise transcriptome assembly.

Finally, normalized reads are assembled with Trinity (Grabherr et al., 2011). I chose this program due to the number of publications done with this procedure, and that was built to assemble transcriptomes, rather than being the adaptation of genome assemblers to this problem, such as Oases (Schulz et al., 2012), Trans-ABySS (Robertson et al., 2010), and SOAPdenovo-trans (Xie et al., 2014). Although Trinity obtains its name because it was composed of three modules, nowadays it is split in four.

First, k-mers are counted with Jellyfish (Marçais & Kingsford, 2011) to estimate and reduce the memory footprint of the procedure. Then, the Inchworm module is executed, where overlapping k-mers are inspected to generate a de Bruijn graph, where nodes are represented by the different k-mers, and links arise whenever two k-mers overlap by k-1 bases. When no ambiguity is possible, k-mers are joined into the initial transcriptomic contigs. Subsequently, Chrysalis clusters the previous contigs by mapping the RNA-Seq reads, connecting contigs into connected components, and revealing alternative spliced transcripts and close gene families. Finally, Butterfly processes every component to report full-length transcripts, isoforms, and separate paralogous genes.

The final object of this procedure is a FASTA file where each record is a transcript, and where the header indicates what gene (connected component) each isoform belongs to.

To check the quality of the assembly, two metrics are computed. For the first, input RNA-Seq reads are mapped back to the assembled transcriptome with Bowtie2 (Langmead & Salzberg, 2012). The reason to do so, is that a well assembled transcriptome should have used all reads in the experiment, and that RNA-Seq reads should not be multi-mapped to different genes. According to the developers, a 90% mapping success rate is a good indicator of a well assembled transcriptome (Haas et al., 2013).

For the second, the transcriptome is searched for conserved Single Copy Orthologs (SCOs) with BUSCO (Simão et al., 2015), to determine how complete the transcriptome is given the number of conserved genes in this database. These SCOs are dependent on the branches of the Tree of Life you are interested in. Therefore, since this work is related to fish, we used the set of Actinopterygii SCOs. This program works by performing homology searches with BLAST (Camacho et al., 2009) and HMMER (Eddy, 2011) to the desired database, and then SCOs are reported as present (single copy or duplicated), fragmented, or missing. The aim of a transcriptome assembly should be to obtain the highest number of SCOs, either complete or fragmented.

Given that the *Clupea pallasii* transcriptome was performed by pyrosequencing, another approach had to be taken. First, reads were corrected and trimmed with PyroBayes (Quinlan et al., 2008), trimmed, and cleaned for contamination with SnoWhite (Dlugosch et al., 2013), and assembled with Newbler (Roche Ltd.), that instead of the de Bruijn approach uses the Overlapping Layout Consensus methodology, based on an all-vs-all sequence alignment to extend reads into transcripts. This sequence of procedures was encapsulated into the *smsk_454* pipeline (Table 4.1).

4.1.4 *smsk_trinotate*: Transcriptome Annotation

Assembled transcripts alone do not possess biological information. The straightforward approach is to first discover CDS and then match them to protein databases

with known annotation, because function is related to protein sequence. We declined studying non-coding RNAs (ncRNAs) because RNA-Seq and genome reference-free based tools are not mature enough in non-model species, especially fishes.

The pipeline used for annotation is the one from TransDecoder (Haas, 2016) and Trinotate (Bryant et al., 2017), but optimized for parallel execution. A schematic representation is shown in Figure 4.3. First, since we are searching for homologies between proteins, we replaced BLAST for Diamond (Buchfink et al., 2015), a program with the same purpose but optimized for speed, between 30 to 50 times faster, and the exact sensitivity. And second, hmmscan is not designed to work in parallel by itself. By dividing the set of CDS and executing them in parallel resulted in a faster execution while multiplying the memory footprint. In the end, the result of the pipeline is a table containing the gene-to-transcript-to-protein relationship along the matches against protein databases and the annotations inherited from them: Gene Ontology terms (The Gene Ontology Consortium, 2019), KEGG pathways (Kanehisa & Goto, 2000) and eggNOG functions (Huerta-Cepas, Szklarczyk, et al., 2016).

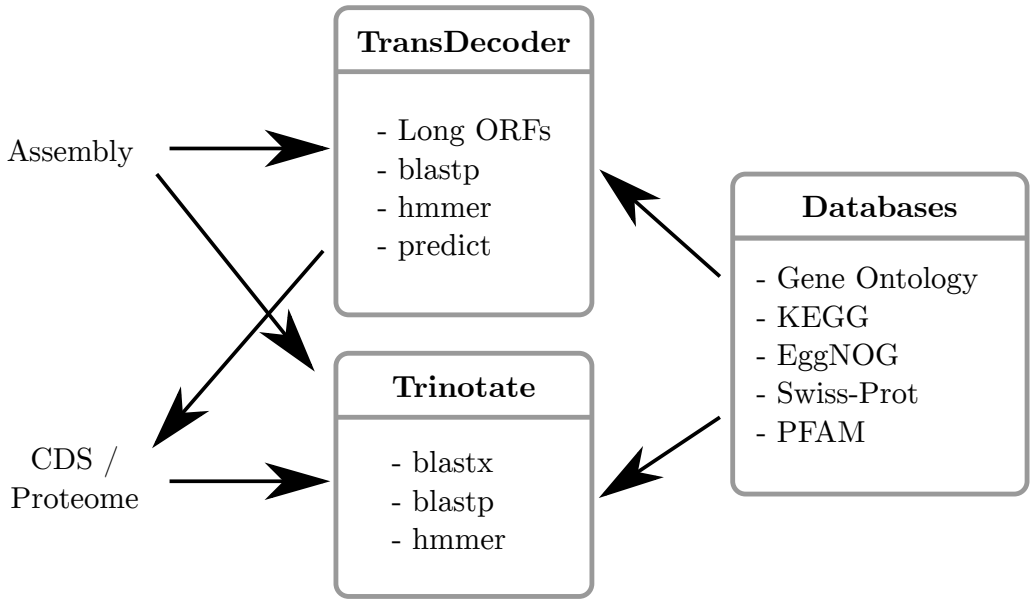


FIGURE 4.3: Schematic representation of the annotation pipeline. Protein-coding sequences are predicted with TransDecoder, which relies also on Blast and HMMER searches. Annotation of the sequences is then derived by a second round of Blast and HMMER searches, and by querying the annotations from UniProt: GO, KEGG and EggNOG.

4.2 Population Genetics of Tench

The purpose of the study of *T. tinca* was to evaluate the presence of the two known phylogroups (Western and Eastern) in two cultured populations in Central Europe: Tabor, from the Czech Republic, and the Hungarian one.

As genetic resources for Tench were very scarce, for this study of population genetics, first, 96 SNP markers were discovered in the transcriptome and, second, were genotyped in 140 individuals. Classical population genetics statistics were applied to genotypic data.

The methods here described are the ones presented in Kumar et al. (2019). The first stages of the procedure consists of the transcriptome assembly and annotation explained in sections 4.1.3 and 4.1.4.

4.2.1 Data sources

For the SNP discovery step, DNA and RNA were parallely extracted. For transcriptome sequencing, the muscle and brain of four individuals were sampled. Two males and two females that pertained to the Hungarian and Tabor breeds cultured in fish tanks at the Faculty of Fisheries and Production of Waters, University of South Bohemia. Additionally, samples were taken under two different metabolic activities, corresponding to winter and summer (at 4 and 20ºC; see Table 4.2). The reason for introducing such diverse conditions (tissue, sex, breed, season) was to maximize the representation of the transcriptome within the sample.

TABLE 4.2: RNA-Seq samples of tench. Experimental design and accession numbers for the RNA-Seq experiment of tench. Two males and two females were used. From them, muscle and brain were extracted. All individuals belong to the Western phylogroup.

Sample	Breed	Season	Sex	Tissue	Sample Name	Accession Numbers
1	Hungarian	Winter	Female	Muscle	MWH-330	SRR6180875
1	Hungarian	Winter	Female	Brain	BWH-330	SRR6180877
2	Hungarian	Winter	Male	Muscle	MWH-389	SRR6180876
2	Hungarian	Winter	Male	Brain	BWH-389	SRR6180878
3	Hungarian	Winter	Female	Muscle	MWH-377	SRR6180879, SRR6180881
4	Tabor	Summer	Male	Muscle	MWT-240	SRR6180880, SRR6180882

Total RNA was isolated using Qiazol lysis reagent (Qiagen). The isolated RNA was quantified with a Nanodrop 2000 (Thermo Scientific) and integrity of RNA (RIN) was determined using an Agilent 2100 Bioanalyzer (Agilent Technologies). Samples with RIN values above 8 were used for RNA sequencing, and used for library construction and Illumina sequencing. According to the RNA quality standards, six samples were sequenced.

For genome sequencing, ten individuals from six locations were collected to maximize genetic diversity, including also the phylogroup of origin (Table 4.3). These samples were obtained from the tissue collection of the Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany, and represent six known populations: Czech Republic, Hungary, Italy, Germany, Turkey, and China.

TABLE 4.3: WGS samples of tench. Ten individuals from 6 different eurasian locations were collected and pooled into two libraries for genome sequencing: Western and Eastern phylogroups.

Sample	Origin	Tissue	Sample Name	Phylogroup	Accession Number
1	Czech Republic	Fin clip	Tench_DNA_1	Western	SRR6180884
2	Czech Republic	Fin clip	Tench_DNA_2	Western	SRR6180884
3	Hungary	Fin clip	Tench_DNA_3	Western	SRR6180884
4	Hungary	Fin clip	Tench_DNA_4	Western	SRR6180884
5	Italy	Fin clip	Tench_DNA_5	Western	SRR6180884
6	Germany	Fin clip	Tench_DNA_6	Eastern	SRR6180883
7	Turkey	Fin clip	Tench_DNA_7	Eastern	SRR6180883
8	Turkey	Muscle	Tench_DNA_8	Eastern	SRR6180883
9	China	Blood	Tench_DNA_9	Eastern	SRR6180883
10	China	Blood	Tench_DNA_10	Eastern	SRR6180883

Genomic DNA was isolated from muscle, fin or blood samples using the peqGOLD Tissue DNA Mini Kit (Peqlab Biotechnologie) following manufacturer instructions. The quantity and quality of DNA was measured with Qubit 2.0 Fluorometer and 0.8% agarose gel electrophoresis. The DNA samples with concentrations $\geq 50\,\text{ng}/\mu\text{l}$, 260/280 ratios of 1.8–2.0 and clear high molecular weight bands on the gel were used for genome sequencing. An equimolar amount of total DNA was then pooled for the library preparation.

In a posterior stage, the SNP genotyping step, 140 tench samples were genotyped from a selected subset of 96 candidate SNPs identified in this study. In all, 66 individuals from the Tabor breed and 74 from the Hungarian breed. In addition, following methodologies previously described in Kocour and Kohlmann (2014), growth hormone (*gh*) gene genotyping was used to classify individuals within the Eastern (E) or Western (W) phylogroup, or the hybrid (H) group. The distribution of samples with respect to the *gh* genotype (E, W, H) are available in Table 4.4. Briefly, nineteen to twenty seven samples were obtained for each phylogroup and breed combination.

TABLE 4.4: Tench samples used for genotyping. The phylogenetic origin is based on genotyping the *gh* gene following Kocour and Kohlmann (2011) and Kocour and Kohlmann (2014).

Breed \Phylogroup	West	Hybrid	East
Hungarian	24	27	23
Tabor	25	22	19

4.2.2 IEB and SNP Calling

Once a reference transcriptome is built, genetic markers can be designed based on it. Nonetheless, a few precautions need to be done in order to successfully genotype them. First, IEBs need to be found. Genotyping chips require each SNP to lie within

an exon so its flanking primers can bind to the DNA sequence. Failing to do so, flanking primers may either be kilobases apart, because each of them will be placed in different exons, or will fail to bind at all because the designed primer will consist of a chimeric sequence made of two exons. If not taken care of soon, these errors will propagate until the stage of large scale genotyping.

Conklin et al. (2013) solved this issue through read mapping (see Figure 3.3). This procedure maps the genomic reads with a read mapper such as Bowtie2 (Langmead & Salzberg, 2012) or BWA-MEM (H. Li, 2013). The procedure deviates from the one in Montes et al. (2013) in two main points. First, the transcriptome is derived from Illumina sequencing, yielding much more reads and less sequencing errors. Second, as the transcriptome is better sampled, it is much more complete and correct, contains a higher number of alternative isoforms per gene, and therefore more processing needs to be done.

Once assembled the transcriptome, given the number of transcripts yielded, and that the number of SNPs to be genotyped (96) was going to be much smaller than the ones discovered (hundreds of thousands), we applied multiple stringent transcript filtering procedures. First, we used the quantification from *kallisto* (Bray et al., 2016) to discard low expressed transcripts. Next, transcripts that according to the annotation were non-coding were removed too. Finally, entire genes composed by two or more coding transcripts were deleted too.

Once the transcriptome was filtered, the parallel mapping approach of DNA and RNA sequences was done. Mappings were done with Bowtie2 with the local and sensitive settings. These alignments were compressed and sorted with Samtools (H. Li et al., 2009). Additionally, possible PCR duplicates were removed with the rmdup subcommand. Finally, SNPs were called with the Samtools mpileup subcommand. A minimum and maximum contig depths 20x and 200x were put in place in order to avoid SNP biases in the lower end, and repetitive sequences and false local alignments in the upper end. For the RNA-Seq to transcriptome mapping, only a minimum contig depth of 8x was required. The remaining SNPs were required to appear at least in 2 RNA-Seq reads, and 3 for WGS reads at the same time. A simplified flowchart of the procedure can be seen in Figure 9.

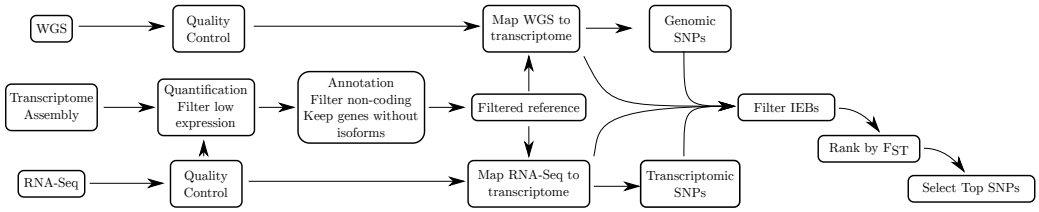


FIGURE 4.4: Flowchart of the IEB detection and SNP calling procedure.

Finally, IEBs were searched in the transcripts with high quality SNPs. The read mappings were reanalyzed in search of locations of sudden ends and starts of alignments. The distribution of such points is compared to a uniform one, and p-values are assigned. Lower p-values mean that there is an excess of starts and ends compared to chance, and therefore indicate the presence of an IEB. Transcripts with signs of IEBs or without match to zebrafish sequences were discarded. Finally, SNPs were ranked by F_{ST} between breeds, and the top 96 SNPs were selected to design probes and primers for the Fluidigm Genotyping System.

For genotyping, 66 Tabor and 74 Hungarian tenches were used (140 in total, see Table 4.4). Then, SNPs were categorized either as no signal (SNP not amplified), disperse (call rate below 80%), monomorphic (minor allele frequency below 1%) and PSV (Paralogous Sequence Variant; all individuals are heterozygotes). To compute the conversion rate (ratio of polymorphic SNPs to genotyped), the no signal and disperse were not taken into account.

Once genotyped, multiple Population Genetics metrics were computed. First, minor allele frequency, expected and observed heterozygosities (H_e and H_o, respectively) were estimated using GeneClass2 (Piry et al., 2004). Second, deviations from the Hardy-Weinberg Equilibrium were obtained through Fisher's exact tests in Genepop (Rousset, 2008), using 10,000 dememorizations, 100 batches and 5,000 iterations per batch. Genetic structure was assessed with Structure 2.3.4 (Pritchard et al., 2000). The number of optimal clusters k was determined with the method proposed in Evanno et al. (2005) by comparing the log-likelihoos of each run with k ranging from 1 to 10. Each clustering was run using a burn-in of 10,000 steps, followed by 100,000 Markov chain Monte Carlo (MCMC) replicates. Barplots were plotted using Pophelper 1.0.7 (Francis, 2017).

Based on the derived population structure, Bayescan 2.1 (Foll & Gaggiotti, 2008) was used to detect loci under natural selection, also known as outlier loci. It was run with 20 pilot runs of 5,000 iterations each, 50,000 burn-in iterations and prior odds of 10 for the neutral model. The False Discovery Rate was corrected using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995). Outlier loci were removed from subsequent analysis, but their annotation was inspected.

Finally, neutral genetic differentiation and inbreeding were measured. Neutral genetic differentiation was estimated with Arlequin 3.5 (Excoffier & Lischer, 2010) by computing the unbiased pairwise F_{ST} (Weir & Cockerham, 1984). Inbreeding was estimated with the F_{IS} statistic using Fstat (Petit et al., 2001). Significance of the F_{ST} and F_{IS} statistics were tested with permutation tests, using 1,000 permutations. False Discovery Rates were corrected this time using the Bonferroni method (Rice, 1989).

4.3 Development of EXFI

As shown in the introduction, I developed an alternative method to IEB, the one used previously in our group to ascertain the population structures of the European anchovy, the Atlantic mackerel, and tench. First I will explain the methodology behind, then an explanation of the datasets used to experiment and why. Finally I show how optimal parameters were found, and how the results were compared to two other tools.

4.3.1 Description of the Algorithm

In a similar approach to the IEB method, we developed EXFI to split a transcriptome into its constituent exons. The procedure is based on locating k-mers in the transcriptome that are not in the WGS experiment because they are made up of two exons.

EXFI is written in Python3, using packages designed for efficient big data processing (Pandas; McKinney, n.d.; Reback et al., 2021) and NumPy; Harris et al., 2020), Bioinformatics (BioPython; Cock et al., 2009), and also high performance tools to manipulate k-mers (BioBloomTools, Chu et al., 2014; ABySS2, Jackman et al., 2017) and genomic intervals (Quinlan & Hall, 2010).

The operation of the procedure can be classified in two steps: k-mer storage and exon prediction, carried out respectively by the *build_baited_bloom_filter* and *build_splice_graph* scripts. An additional script is provided to convert the splice graph, in GFA1 format, into FASTA format, either as separated exons or gapped transcripts.

The key data structure for EXFI is the Bloom filter, a probabilistic data structure used in this case to store k-mers fast and efficiently. It cannot produce false negatives, but false positives, although the rate they are produced can be controlled.

The method to process and store k-mers is to use hash functions. Hash functions map uniformly the k-mers to a range of integer numbers, in a way that similar k-mers are sent to very different numbers. The problem of this approach is that collisions can happen: two different k-mers can have the exact same hash value, and therefore cannot be distinguished. To deal with this problem, we can, instead of using a single function, use multiple independent hash functions. Therefore, the probability of two different k-mers having the exact same values decreases dramatically. Another desired property of hash functions is that they need to be extremely fast.

The recipe for a Bloom filter are hash functions and a bit table. Every time we want to remember a k-mer, we only need to store a one in the positions that the hash values give, and to check if a k-mer has been seen, we only need to check if all positions given by the hash value are set to one. The drawback of this approach is that as the number of inserted elements grows, the probability of reporting a false positive

k-mer grows too, since the hash value of a k-mer could be colliding with the hash values of multiple k-mers at the same time.

Therefore Bloom filters have a False Positive Rate (BF FPR) and no False Negative Rate. This means that it is possible for the data structure to report k-mers that have never been seen in the experiment. This BF FPR, the size of the bit table, the number of hash functions used, and the number of different k-mers are the four parameters and are related one to each other. For example, lower BF FPR can be achieved by either increasing the table size, increasing or decreasing the number of hash functions, or decreasing the WGS experiment.

Since the transcriptome is a fraction of the genome, a very significant fraction of the WGS experiment is not necessary to discover exons in the transcriptome. Therefore, prefiltering of the WGS experiment with an additional BF can significantly reduce the BF FPR. For example, the human exomes suppose around 1% of the genomic sequence, and therefore this filtering can reduce the number of sequences to be processed by an order of magnitude, and therefore either decrease the BF FPR, or reduce the table size.

Finally, every single sequencing error is propagated into k k-mers, all of them unique and therefore will occupy much space in the Bloom Filter. To minimize their impact, we can use Cascading Bloom filters (Salikhov et al., 2014), a similar data structure that uses two or more bit tables instead of one. The way this works is by trying to insert a k-mer in the first table. If it is already in the first table, try to insert it in the second, and so on. Therefore, in the first table you have the k-mers that have at least a frequency of one, in the second table the k-mers that have a frequency of at least two, and so on. EXFI uses only two tables, and by keeping only the second, we can discard the vast majority of sequencing errors and therefore reduce the BF FPR.

Once computed the BF, the second part of the procedure extracts the exons from the transcriptome. To do so, each transcriptomic k-mer is queried in the genomic BF (Figure 4.5a). Whenever a k-mer is missing it must be because an exon is ending and another one is starting at that place. Found k-mers are assembled and further inspected (Figure 4.5b).

The False Positives introduced by the BF have two effects in the exon prediction part: the appearance of very small exons of length, k overlapping the larger ones, and also the extra bases appended to the borders of the exons. To prevent the first, a filtering step is done to remove small exons (length k+5 by default; Figure 4.5c). Then, exons that overlap each other by an excessive number of bases (10) are merged together: the probability of 10 bases overlapping by chance is FPR^{10} (if FPR = 1%, that probability is 10^{-11}, astronomically low; Figure 4.5d). Finally, the remaining overlaps are polished by inspecting the donor/acceptor signal (GU/AG) at the sequence level to finally separate the exons (Figure 4.5e). The result of this step is a

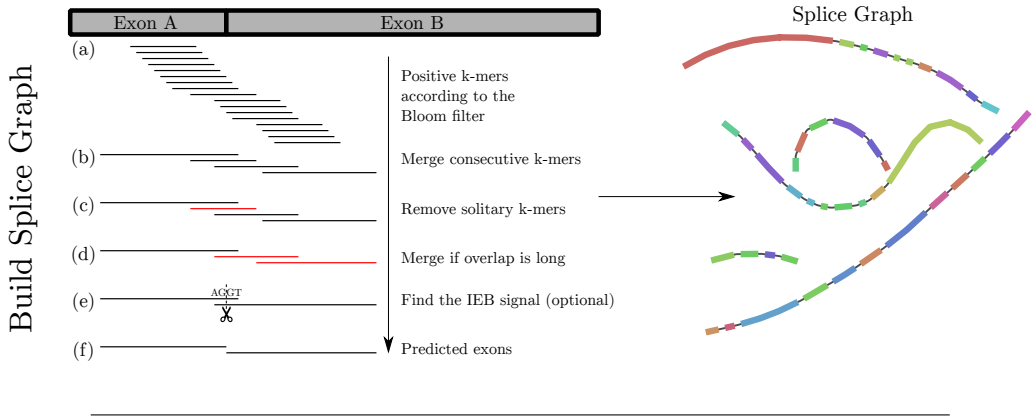


FIGURE 4.5: Flowchart of the EXFI procedure. (a) k-mers are queried sequentially and (b) merged if they overlap by k-1 bases. (c) Solitary k-mers are removed as they are false positives. (d) Merge long overlaps. (e) Polish the splice graph by finding the donor and acceptor splice sites. Figure adapted from Langa et al. (2020)

GFA1 file, a common format to represent assembled genomes as linked contigs, and that can be visualized with Bandage (Wick et al., 2015).

Finally, the *gfa_to_fasta* script is provided to either extract the individual exons or the spliced transcript (a sequence in which exons are separated by regions filled with N letters) in FASTA format. These files can be used in downstream analyses, such as the ones done for SNP discovery in section 1.4.2.2.

Therefore EXFI provides user-friendly tools for researchers working in non-model species and would like to study the population genetic variation associated with coding sequences and/or the exome.

4.3.2 Datasets Used

Since genomes and transcriptomes vary widely across the Tree of Life, and because the end in mind of EXFI is to discover variants for population studies, we defined three tiers of datasets to prove the concept: references, semi-references, and mega-genomes without a reference.

For the first tier, we wanted to test EXFI in a very controlled scenario by applying it to two well known genomes. We used the zebrafish and human genome and transcriptome references (GRC38 and GRCz10, Ensembl release 91, December 2017). Instead of using real WGS datasets, multiple samples were generated with *wgsim* (H. Li, 2011), simulating each individual by adding different variants and sequencing errors through different random seeds. These two datasets were the ones that guided the development of EXFI.

As the human and zebrafish are far from the real-world application of EXFI, I moved a little further experimenting with reference transcriptomes but real WGS datasets. For the second tier, I obtained WGS reads from ENA for the Atlantic salmon (*Salmo*

salar; Kijas et al., 2018) and the Atlantic herring (*Clupea harengus*; Lamichhaney et al., 2012). Instead of relying on individually sequenced samples, Pool-Seq libraries (sequencing multiple individuals without a differentiating tag) were used to measure the robustness of the predictions under the influence of tens to hundreds of haploid genomes in the same dataset. Moreover, a salmonid-specific WGD occured 80 mya, and according to Ensembl, the salmon genome is double in size compared to zebrafish (3.41 Gbp vs. 1.37 Gbp; 47,329 vs. 25,592 protein-coding genes; Ensembl Release 104), complicating the procedure even more. Additionally, for the Atlantic herring we used both the available reference transcriptome along with an assembled one, to compare the effects of a *de novo* assembly in exon prediction.

Finally, to test the upper limit of EXFI, we included a third tier of species composed of the axolotl (*Ambystoma mexicanum*; Keinath et al., 2015) and the sugar pine (*Pinus lambertiana*; Neale et al., 2014), the first known for its regenerative capacity and the second for being the largest species of pine. Specifically, their genomes are also known to be massive, 32 and 28 Gbp in size, respectively.

Additionally, since the motivation for this objective was to improve the SNP discovery procedure done in Objective 1, we reanalyzed the *Tinca tinca* transcriptome and WGS effort, for which no reference genome assembly is available to compare the results. Table 4.5 resumes the datasets used, and shows the references and experiments used.

Therefore, we tested EXFI under a wide variety of reference status, genome sizes, and experimental WGS efforts.

TABLE 4.5: Datasets used for the development, validation and application of EXFI.

Experiment	Zebrafish	Human	Atl. salmon	Atl. herring	Sugar pine	Axolotl	Tench
Genome Type	Chromosome	Chromosome	Chromosome	Scaffold	Scaffold	Chromosome	Not Available
Genome Size (Gbp)	1.34	3.09	2.97	0.81	27.60	32.40	0.78
Genes	25,497	21,407	79,030	25,135	Unknown	Unknown	Unknown
Transcriptome Type	Reference	Reference	Reference	Reference / *de novo*	*de novo*	*de novo*	*de novo*
Transcripts	51,714	164,776	109,584	29,353 / 97,777	331,11	180,605	267,058
Transcriptome Size (Mbp)	110.69	270.48	355.21	64.18 / 55.39	36.74	229.48	294.70
Exons	495,200	1,199,596	1,313,909	314,220 / Unknown	Unknown	Unknown	Unknown
Samples	2	6	20	50	1	1	10
Reads (M)	720.00	2,160.00	1,259.27	418.73	9,300.90	7,121.91	318.72
Total Bases (Gbp)	72.00	216.00	125.93	41.13	1395.13	712.19	31.87
Coverage	53.73	69.90	42.44	50.92	50.54	21.98	51.58
Genome source	GRCv10	GRC38	GCA_-000233375.4	GCA_-000966335.1	GCA_-001447015.2	GCA_-002915635.2	NA
Transcriptome source	GRCv10	GRC38	GCA_-000233375.4	GCA_-000966335.1 / SRR611605	GEUZ01	GFZP01	GFZX01
WGS source	Simulated	Simulated	ERR2247296-ERR2247299	SRR611633-SRR611635	SRR2026990, SRR2026995, SRR2026998, SRR2027002-SRR2027013, SRR2027088-SRR2027101	SRP051662	ERR1725872, ERR1725873

4.3.3 Validation

To validate and compare the results, we provide the *compare_to_gff3* script to compare the predicted exons to a known reference in GFF format. Comparisons are given in terms of the usual classification metrics:

- True Positives (TP), correctly predicted exons;
- False Positives (FP), incorrectly predicted exons;
- False Negatives (FN), missing exons;
- Precision (P), the ratio of correct predictions, P = TP / (TP + FP);
- Recall (R), the ratio of predictions found: R = TP / (TP + FN); and
- F_1 score, the harmonic mean between precision and recall: F1 = 2 PR / (P + R).

These measurements were complemented with the mapping of the predicted exons to the reference genome, when available. This was done with BWA-MEM (H. Li, 2013) and Samtools (H. Li et al., 2009). Statistics are given by the number of mapped exons, and how many of those were done with a perfect CIGAR string (100% matched, allowing only with small insertions or deletions, but no base clipping).

For the validation of the tool, the datasets described in section 4.3.2 were used. Briefly, zebrafish and human datasets were used because they are species deeply studied. We moved a little outside reference territory by using Atlantic salmon and Atlantic herring, both fishes, the central theme of this thesis. The first was chosen because it has a recent genome duplication, and the second to assess the differences between assembled and reference transcriptomes, and for being a clupeid fish. Additionally, both WGS experiments came also from a Pool-Seq approach, rich in variants but also in sequencing errors, and therefore should misguide the exon prediction process by increasing the BF FPR. The last two species used were the axolotl and the sugar pine, known for having mega-genomes and their repetitive content.

First of all, it was necessary to discover the optimal four parameters that impact the performance of EXFI. The annotation-based metrics are the ones used to discover them. These four parameters are: to filter or not the WGS experiment, the quantity of memory used, the k-mer length, and the coverage of the genome used. They were discovered over the zebrafish and human datasets, since they are the most complete references.

The effect of filtering or not reads was the first measured. Non-exonic reads and their sequencing errors will fill the Bloom filter unnecessarily, but the filtering step itself will increase the runtime. To measure the differences, EXFI was applied with and without the read filtering step, with the k-mer length fixed to 25.

Low memory footprint and high speed are the key reasons to use Bloom filters, but using as little memory as possible is our target. The effect of decreasing the memory requirements results in the increase of the BF FPR and therefore a decrease of Precision and Recall of the results. The trade-off between memory and BF FPR was done by repeatedly analyzing the zebrafish datased, with BF sizes from 4 to 60 GB in steps of 4 GB, and fixing the k-mer length to 25 base pairs.

The third parameter, the k-mer length, was analyzed using the odd values in the range from 21 to 65 base pairs, while also using Bloom filters of 4 and 60 GB in size, over the zebrafish dataset only. The effect of the k-mer length is very sensitive. If k is too low, a k-mer can be present in too many regions at the same time and become too unspecific, and therefore more reads (and more sequencing errors) are inserted into the data structure, and the BF FPR will increase, therefore decreasing the Precision. If k is too high, there will be less elements inserted and with less frequency, and they will be susceptible to be eliminated by the cascading Bloom filter, lowering the BF FPR but at the same time the Recall of the method. Also, the method is myopic: it cannot see exons of length k or less, and therefore more false negatives will be reported if k is too high. Therefore, a balanced k-mer length needs to be found.

Finally, a correct genome coverage needs to be used. If the WGS experiment is too shallow, the method will overpredict IEBs because k-mers will be missing everywhere. On the opposite side, if the sampling is too deep, the number of sequencing errors will become so endemic that the frequency filter will start to accept sequencing errors as truthful k-mers. For this case, the zebrafish dataset was sampled in 10% increments with *seqtk* (H. Li, 2012), and each subsample was analyzed using the high and low memory settings and k-mer length as before.

The performance of EXFI was also compared to two other tools: ChopStitch (Khan et al., 2018) and GMAP (Wu & Watanabe, 2005) using the metrics described above, and also BF FPRs, speed and memory footprint. ChopStitch is a tool similar to EXFI that uses a transcriptome, WGS reads and Bloom filters to predict the exome. The main difference is that in this tool the BF FPR needs to be provided first, and then the entire WGS dataset is processed to estimate the needed memory and hash functions. Additionally, it uses the entire WGS dataset to predict exons. The second one, GMAP, is a tool designed to perform gapped alignments of Expressed Sequence Tags (ESTs) and full transcripts to a reference genome. The six datasets (zebrafish, human, salmon, herring, pine and axolotl, see Table 4.5) were used to compare the three methods in terms of the metrics described above: annotation and mapping-based.

To give an example of the power of EXFI over a non-model species, we reanalyzed the tench dataset, not only splicing the transcriptome but also to discover SNPs, specially the 96 used for genotyping. The steps done to find SNPs were the use of EXFI itself, read mapping with Bowtie2 (Langmead & Salzberg, 2012) and variant calling with BCFTools (H. Li et al., 2009). Variants with quality value below 20

were discarded, along those that were 35 base pairs or less to other variants or exon boundaries.

4.4 Towards a Complete Transcriptome for *S. pilchardus*

This section describes the strategy and methodology behind the sequencing effort of the European sardine.

4.4.1 An Optimal Sampling Strategy

Extracting RNA from tissues is a very complicated and delicate procedure. Given the throughput of current sequencing technologies, RNA-Seq provides the expression of thousands of genes. Nonetheless, transcript expression is dependent on the tissues used, and the expression of a handful of transcripts can eclipse the entire experiment. Therefore, I wanted to find how much a transcriptome changes when multiple tissues are used, and what could be a cost-effective strategy.

Before carrying out any experiment in the lab, I did a simulation over the transcriptomes that we would obtain by subsampling twelve libraries from zebrafish. We looked towards the zebrafish because of the enormous genomic resources available. Concretely, we used the zebrafish dataset from Pasquier et al. (2016). In this paper, the Actinopterygii lineage is sampled to construct and quantify 27 transcriptome assemblies derived from twelve tissues. The RNA-Seq dataset is available under accession number SRP044781 at ENA. Details are available in Table 4.6.

TABLE 4.6: Zebrafish samples for optimal strategy discovery. The twelve libraries were subsampled to obtain an optimal sequencing strategy to either sample deeply a tissue, or shallowly use all of them. The number of Paired-End reads is expressed in millions.

Library	Accession number	Reads (M)
Bones	SRR1524244	96.72
Brain	SRR1524238	35.36
Embryo	SRR1524246	55.19
Gills	SRR1524239	54.47
Heart	SRR1524240	85.67
Intestine	SRR1524245	43.19
Kidney	SRR1524243	46.37
Liver	SRR1524242	59.25
Muscle	SRR1524241	34.03
Ovary	SRR1524248	22.03
Testis	SRR1524249	59.90
Unfertilized eggs	SRR1524247	24.88

To find out a near-optimal, simulations on this *D. rerio* RNA-Seq dataset were performed. For every tissue, the dataset was subsampled incrementally from 1 to 20

million paired-end reads, in increments of 1 million using *wgsim* (H. Li, 2011). Additionally, a mix of the 12 tissues was sampled in the same way to observe the capabilities of a varied sequencing effort, using the same number of reads.

Then, for every one of the 260 library and subsample combinations, transcriptomes were assembled both *de novo* and reference-based with Trinity (Grabherr et al., 2011) and StringTie (Pertea et al., 2015). In the *de novo* case, transcripts were associated with its gene using BLASTN (Camacho et al., 2009), while in the reference case, transcript identifiers were extracted from the resulting GFF3 files. Then, we proceeded to analyze the number of genes yielded by the twelve tissues and the mix, either when a genome assembly is present or not.

4.4.2 Sequencing of *Sardina Pilchardus*

Given the results obtained for zebrafish, we decided that, when the sequencing effort is limited, the optimal strategy to obtain the best representation of the transcriptome of a given species is to use all possible tissues. Additionally, given that the ultimate use of this dataset is to support a future SNP discovery procedure, to minimize individual variation that misleads the transcriptome assembler, the number of samples involved had to be reduced to the minimum.

Three individuals from the European Atlantic Ocean were collected by the IFREMER institute during the EVHOE scientific surveys (October 10th, 2015; Leaute et al., 2015). From these individuals, nine tissues (brain, eye, heart, kidney, liver, muscle, ovaries, skin, and testes) were dissected onboard, immediately immersed in RNAlater (Invitrogen), and stored at -20ºC until further processing. One of the individuals was used as the main donor of tissues, a female, while another individual was used as a testes donor, and the third one to complete the sampling, a kidney and skin donor.

Total RNA was extracted using TriZol®Reagent (Life Technologies) and quantified with Agilent 2100 Bioanalyzer combined with Agilent RNA 6000 Nano chips (Agilent Technologies, Inc.) at the Gene Expression Unit (SGIker) of the University of the Basque Country UPV/EHU. Samples with RIN below 8 were immediately discarded. For every tissue, the sample with the highest RIN was used for sequencing. The exceptions were testes, since there was only one male specimen, and ovary, where both samples were used. A multiplex sequencing library was prepared by labeling each sample with specific 10-mer barcoding oligonucleotides. The barcoded RNA-Seq libraries were sequenced using the Illumina HiSeq 2000 platform using one single lane. Sequencing reactions were performed with paired-end 101-bp and strand-specific protocol at the sequencing facility of the CNAG (Centre Nacional d'Anàlisi Genòmica, Barcelona, Spain). Base-calling was performed using the Illumina native software. Details of the RNA-Seq libraries are available in Table 4.7.

TABLE 4.7: Transcriptome sequencing of *Sardina pilchardus*. Ten tissues were used in total, coming from three individuals. The number of reads is expressed in millions.

Library	Sample	Accession number	Reads (M)
Brain	Female 1	ERR5925802	6.11
Eye	Female 1	ERR5925803	5.34
Heart	Female 1	ERR5925804	4.98
Kidney	Male	ERR5925805	6.68
Liver	Female 1	ERR5925806	4.67
Muscle	Female 1	ERR5925807	5.31
Ovary 1	Female 1	ERR5925808	6.64
Ovary 2	Female 2	ERR5925809	6.57
Skin	Male	ERR5925810	5.17
Testes	Male	ERR5925811	5.04

RNA-Seq reads were processed using the *smsk_khmer_trinity* pipeline described in Section 4.1.3. Briefly, reads were trimmed with Trimmomatic, normalized with khmer, and assembled with Trinity. Quality of the assembly was measured in terms of mappability of the trimmed RNA-Seq reads with Bowtie2 and the number of Actinopterygii SCOs present. The annotation was obtained by predicting protein-coding genes with TransDecoder and annotations were derived with Trinotate, using the *smsk_trinotate* pipeline from Section 4.1.4. Trimmed reads were quantified with *kallisto* (Bray et al., 2016)) and normalization in terms of Transcripts Per Million (TPM) was obtained with *sleuth* (Pimentel et al., 2017).

This study has defined an optimal strategy for the construction of the transcriptome of a non-model species, its annotation and specific tissue expression of the European sardine, a species with a growing interest from the genetics perspective.

4.5 Clupeiformes and Genes under Positive Selection

This section describes the procedure behind the discovery of genes under positive selection in Clupeiformes. It is divided into the species sampling, the transcript clustering into genes, the construction of the species tree, and the discovery of positively selected genes. We followed the methods from Roux et al. (2014) and Ciezarek et al. (2016), which in turn follow the procedures from Ensembl Compara (Herrero et al., 2016) and Selectome (Moretti et al., 2014; Proux et al., 2009). Given the complexity and scale of the analyses required, the computational steps were packaged into a Snakemake pipeline: *smsk_selection*. A schematic representation of the procedure is available in Figure 4.6.

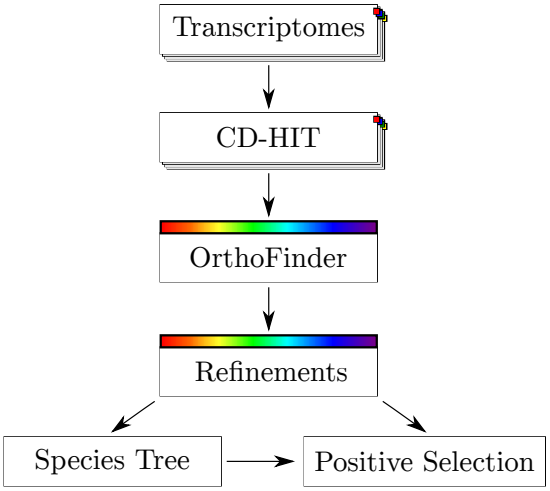


FIGURE 4.6: Flowchart of the *smsk_selection* pipeline. Coding sections of the transcriptomes are clustered with CD-HIT, and then clustered into orthogroups with OrthoFinder. Multiple rounds of orthogroup and sequence refinements before constructing the species tree. Finally, genes under positive selection are identified.

4.5.1 Sampling

Multiple data sources were considered to study the evolution of Clupeiformes (see Table 4.8).

First, the transcriptomes from *Clupea harengus* (í Kongsstovu et al., 2019) and *Denticeps clupeoides* (Rhie et al., 2021) were obtained from the NCBI Genome resource. These transcriptomes are automatically annotated by predicting genes and transcripts *ab initio* and by mapping related RNA-Seq experiments.

Second, previously assembled and published transcriptomes were obtained for *Konosirus punctatus* and *Alosa pseudoharengus* (Czesny et al., 2012), both available at the Transcriptome Shotgun Assembly database from the European Nucleotide Archive.

Third, *de novo* transcriptome assembly was performed for species *Engraulis encrasicolus*, *Coilia nasus* (Zhu et al., 2017), *Clupea pallasii* (Roberts et al., 2012), *Tenualosa ilisha* (Divya et al., 2019), *Brevoortia tyrannus* (Iv et al., 2017), *Alosa alosa* (Pasquier et al., 2016), *Sardina pilchardus* (Langa et al., 2021), and *Sardinops sagax* (Richards et al., 2018). To do so, raw RNA-Seq experiments were downloaded from ENA. Transcriptomes were assembled using the methods described in section 1.2.3. Finally, to help with the gene and phylogenetic clustering procedure, additional transcriptomes were downloaded from Ensembl for the fish species *Astyanax mexicanus*, *Danio rerio*, *Gasterosteus aculeatus*, *Gadus morhua*, *Latimeria chalumnae*, *Lepisosteus oculatus*, *Oryzias latipes*, *Oreochromis niloticus*, *Poecilia formosa*, and *Xiphophorus maculatus*; and mammals *Homo sapiens* and *Mus musculus* (Ensembl release 98).

TABLE 4.8: Clupeiformes species studied. Species, tissues, sequenced bases and accession numbers of the species used in the study of Clupeiformes. Gbp: giga base-pairs.

Species	Source	Gbp	Transcripts	Accession number	Tissues
D. clupeoides	NCBI Annotation	-	59,645	GCF_900700375.1	-
E. encrasicolus	Illumina HiSeq 2000	15	719,059	PRJNA261165, PRJNA348159	Juveniles, Kidney, Liver, Muscle, Ovary, Testes.
C. nasus	Illumina GAII, HiSeq 2000, HiSeq 2500	55	885,281	PRJNA235378, PRJNA242212, PRJNA251948, PRJNA315471	Liver
C. pallasii	Roche 454 GS FLX Titanium	0.6	19,004	SRX022719, SRX082042	Liver, Testes
C. harengus	NCBI Annotation	-	46,203	GCF_900700415.1	-
K. punctatus	ENA TSA	-	69,974	GHHF01000000.1	Muscle, Liver, Gill, Heart, Kidney, Swim Bladder, Sexual Gland
T. ilisha	Illumina HiSeq 2500	6	107,804	PRJNA396091	Liver
B. tyrannus	Illumina HiSeq 2000	18	266,785	PRJNA276563, PRJNA281114	Testes
A. alosa	Illumina HiSeq 2000	66	734,830	PRJNA256955	Bones, Brain, Gills, Heart, Intestine, Kidney, Liver, Muscle, Ovary, Testes.
A. pseudoharengus	ENA TSA	-	216,529	GFCK01000024	Gill
S. pilchardus	Illumina HiSeq 2000	6	198,597	PRJEB18441	Brain, Eye, Heart, Kidney, Liver, Muscle, Ovary, Skin, Testes.
S. sagax	Illumina GAIIx	5	196,984	SRR7955963	Liver

Since we are only interested in protein coding sequences, they were extracted using TransDecoder following the *smsk_trinotate* pipeline (see Section 4.1.4).

The remainder of this section explains the *smsk_selection* pipeline (Table 4.1). Briefly, it consists of three steps: 1) the clustering of all the species transcripts into gene families, 2) the construction of the species tree, and 3) the detection of positive selection of genes and functional groups of them.

4.5.2 Clustering

The first step involves the clustering of the transcripts into genes. There are four problems when trying to group transcripts of multiple species into a single gene cluster. The first is alternative splicing, in which a single gene produces multiple transcripts. The second is the Teleost-Specific WGD, which may misguide the clustering step by merging duplicated genes present in all species into a single cluster. The third is that evolution shapes gene evolution in different manners, modifying, deleting and duplicating genes when necessary. The fourth problem is that gene expression is tissue-specific, and therefore it is not possible to discern between unexpressed and deleted genes.

First, protein-coding sequences for the 24 species (12 clupeid and 12 non-clupeid) were clustered separately with CD-HIT-EST (W. Li & Godzik, 2006) with a similarity threshold of 99.5% to remove within-species redundancy. Then, all non-redundant CDS were clustered together with OrthoFinder (Emms & Kelly, 2019) to predict orthogroups (clusters of homologous sequences - putative genes or gene families) across all species. This program first performs an all-vs-all Diamond search (Buchfink et al., 2015) of the translated CDS sequences, then normalizes the results through their bit scores and taking into account sequence lengths and phylogenetic distances, and then performed clustering with MCL (van Dongen & Abreu-Goodger, 2012). Subsequent steps on the pipeline include the inference of the species tree (Emms & Kelly, 2017), and the separation of the orthogroups into orthologs and their phylogenetic trees associated.

An additional separation and refinement of the orthogroups is done through the methods presented in Y. Yang and Smith (2014). It consists of two rounds of protein realignment with MAFFT (Katoh & Standley, 2013), column trimming with pxclsq (Brown et al., 2017), tree inference with RAXML-NG (Kozlov et al., 2019), using the WAG protein evolution model (Whelan & Goldman, 2001), trimming of tips of the trees that had an absolute length of 2 or a relative length of 10 times its sister tip (*trim_tips.py*), removal of tips from the same species with fewer characters while also removing the paraphyletic ones (*mask_tips_by_taxonID_transcripts.py*), and removal of deep paralogs (tips with a branch length greater than 0.5; *cut_long_internal_branches.py*).

Then, high-quality orthologs were predicted using the Root-to-Tip method (*prune_paralogs_RT.py*), which takes into account gene duplication events, in particular the Teleost-Specific WGD. Then, non-clupeid species were used as outgroups and removed from all the clusters.

Given that alignment errors are an important source of false positives when searching for Positive Selection (Löytynoja, 2014; Redelings, 2014), a more stringent procedure was applied to align CDS and proteins, made up also of two rounds of alignment and refinements. First, proteins were realigned with M-Coffee (Wallace et al.,

2006), which aligned independently each orthogroup with Muscle (Edgar, 2004), MAFFT (Katoh & Standley, 2013), T-Coffee (Notredame et al., 2000), and kalign (Lassmann et al., 2009). These alignments were then evaluated, and columns with a score below 9 (out of 9) were removed. Finally, proteins were back-translated to CDS, where columns with occupancy below 50% were removed with pxclsq, and rows rich in gaps were removed too with MaxAlign 1.1 (Gouveia-Oliveira et al., 2007) using default settings. Clusters were reprocessed again, and were the ones used to build the species tree and to search for positive selection.

4.5.3 Species Tree Construction

Due to the redundancy of the genetic code, multiple codons encode the same amino acid. Four-fold degenerate sites are the positions in the third base of codon alignments that produce the same amino acid no matter what mutation occurs in that position. Therefore, a phylogeny built on these positions minimizes the effects of positive selection over the species tree (Eyre-Walker & Keightley, 1999; Nachman & Crowell, 2000). Four-fold degenerate sites were extracted from all the orthogroups previously aligned, and were concatenated into a supermatrix, requiring each orthogroup to contain at least four taxa. Then, columns were removed if they had occupancy of less than 50% with pxclsq, and converted to PHYLIP format. ModelTest-NG (Darriba et al., 2020) was used to estimate the optimal evolution model to construct the phylogenetic tree. Then, RAxML-NG was used to obtain the Maximum Likelihood (ML) tree, performing 1,000 bootstrap replicates. Finally, ExaBayes 1.5 (Aberer et al., 2014) was run to obtain the Bayesian phylogenetic species tree, using the ML tree as the starting tree, four independent MCMC runs, 3 coupled chains, and one million generations each, and sampling every 500 generations. The sdsf and postProcParams programs were run to ensure that the split frequencies, scale reduction factors, and effective sample sizes were close to zero, one, and 200, respectively. Finally, using the bootstrap replicates, a consensus unrooted tree was generated, and from this one, a rooted version was generated with *pxrr* using *D. clupeoides* as outgroup.

4.5.4 Finding Signals of Positive Selection

To find genes under Positive Selection, we used the branch-site test (Zhang et al., 2005). This method studies the ratio between nonsynonymous (dN) to synonymous (dS) substitutions, denoted by ω (ω = dN/dS), in the branch of interest (the foreground branch) compared to the remainder (the background). On the one hand, the null model assumes that the evolution speed of the foreground branch is strictly less than one ($\omega_f < 1$), i.e., there is conservation in this branch, no matter what is happening in the background. On the other hand, the alternative hypothesis allows four situations:

- Codons are conserved in both branches: $0 < \omega_f, \omega_b < 1$,

- Codons are evolving neutrally: $\omega_f = \omega_b = 1$,
- The foreground branch is under positive selection, while the background is evolving neutrally: $\omega_f > 1 = \omega_b$, and
- The foreground branch is selected, while the background is fixed: $\omega f > 1 > \omega_b$.

For each branch and ortholog analyzed, a likelihood ratio test (LRT) is used to compare whether the alternate model fits better than the null hypothesis by comparing it to a χ^2 test.

The go-to program to compute these tests is CodeML, part of the PAML's package (Z. Yang, 2007), but instead we used ETE3 (Huerta-Cepas, Serra, et al., 2016) for its convenience: it acts as an interface through the command line, avoiding designing one control file per branch and orthogroup, parsing the very verbose outputs of CodeML, and being able to plot each alignment, which contains also the p-values and ω ratios for the different models used. Under ETE3, the models used are denoted bsA1 (the null hypothesis, branch-sites are under relaxation), and bsA (the alternative hypothesis, branch-sites are under positive selection). Prior to testing for selection, each orthogroup had to contain at least two species both in the foreground and background branches.

Due to the sensitivity of the method to initial conditions (Z. Yang & dos Reis, 2011), for every ortholog and branch, this LRT needs to be executed multiple times with different starting values for ω_0 (0.5, 1.0, and 1.5). Orthologs and branches were considered putatively under selection if the tests were found significant (p-value < 0.05) in the three different starting points.

Since alignment error is a key source of false-positives (Markova-Raina & Petrov, 2011; Redelings, 2014), putatively selected orthogroups were processed again with Guidance2 (Sela et al., 2015), a codon-aware probabilistic aligner known to produce low rates of false positives compared to previous approaches. This program was executed using PRANK (Löytynoja, 2014), performing 100 bootstrap alignments. Low-quality positions in the alignments were removed this time with TrimAl 1.2 (Capella-Gutiérrez et al., 2009), using the automated feature, and taxa rich in gaps were removed too with MaxAlign. Finally, a second step of detection of positive selection was performed, this time with FastCodeML 1.3.0 (Valle et al., 2014), using also the different starting ω_0. As in ETE3, both foreground and background branches had the requirement of a minimum of two species.

Mitochondria were analyzed and processed similarly, but separately since they have their own genetic code. The 13 mitochondrial genes for all clupeoid species but one (*S. sagax*) were downloaded from NCBI Gene and separated into 13 different FASTA files. They were directly aligned with GUIDANCE2, trimmed with TrimAl and filtered with MaxAlign, and searched for positive selection with ETE3 three

times, one per starting point, using this time the vertebrate mitochondrial genetic code.

All p-values from the two methods (ETE3 and FastCodeML) and three starting points were merged across all orthologs and all branches analyzed, and corrected with the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). All ortholog-branches with all corrected p-values under 0.05 were considered to be under positive selection.

Since synonymous-site saturation may influence the reliability of the branch-site test (Gharib & Robinson-Rechavi, 2013; Roux et al., 2014), CodeML's free-ratio (b_free) and one-ratio models (M0) were run again with ETE3 in the selected orthogroups-branches. Orthogroup-branches with dS > 1 in both models were considered to be under synonymous-site saturation, and therefore possible false positives.

Then, every high-quality orthogroup was annotated. Every transcript was searched for homology against the Ensembl's zebrafish transcriptome with DIAMOND's BLASTX implementation, matching each clupeid transcript to a single zebrafish transcript, and therefore a zebrafish gene via the lowest e-value. Then, each orthogroup was matched to a single Ensembl gene identifier and gene symbol by a majority vote of its constituent transcripts. If multiple orthogroups matched to the same gene symbol, possibly due to a gene duplication event, a suffix of the form "-n|m" was added, indicating that such an orthogroup is the copy number n out of m. Orthogroups that did not match any zebrafish gene were simply given the symbol "unknown" and added also a suffix. Tables with the equivalences between the zebrafish genes and their Gene Ontology (The Gene Ontology Consortium, 2019) and Reactome (Jassal et al., 2020) annotations were downloaded from Ensembl's Biomart (Kinsella et al., 2011). Then, each orthogroup inherited the annotations of its corresponding zebrafish gene. Additionally, gene family annotations and human to zebrafish ortholog equivalences were downloaded from the HGNC database (Braschi et al., 2019) and Ensembl's Biomart. Each clupeoid orthogroup inherited the gene family annotation of their zebrafish ortholog, which in turn inherited it from its human equivalent. When possible, HGNC families were subdivided into subfamilies for a more granular analysis. For example, the human "Aldo-keto reductase family" could be subdivided into subfamilies "Aldo-reductase family 1" to "Aldo-keto reductase family 7". This division of the gene families can discover subfamilies under positive selection while its superfamily is not.

Enrichment of Gene Ontology, Reactome, and HGNC clusters was analyzed with the Bioconductor (Huber et al., 2015) ClusterProfiler package (Yu et al., 2012). This R (R Core Team, 2020) package performs enrichment of any custom set of terms with the *enrich* function and provides multiple forms to visualize the results. For the three collections of terms, we used as the foreground those orthogroups that were selected and had an annotation, and as the background list the set of orthogroups that were

annotated. In both cases, terms associated with between 5 and 500 terms (minGsSize = 5, maxGsSize = 500) were kept to avoid very specific and non-specific categories, respectively. For each term, a Fisher exact test was performed (Fisher, 1934), and all p-values were corrected by the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). Additional data processing and visualization were done with the R Tidyverse metapackage (Wickham, 2017).

Finally, we tracked the number of times an orthogroup was positively selected, to obtain which ones are under strong evolutionary pressure, putatively indicating not only positive selection but also divergence.

This study demonstrates that selection has acted, among many other things, on the transport and storage of lipids in Clupeiformes. These fishes are already known for being a rich source of protein and lipids, particularly LC-PUFAs of the ω-3 kind, of great importance for their protective health benefits.

Chapter 5

Summary and Discussion of the Results

The main summary of this Thesis is that RNA-Seq datasets are very powerful, both in microevolutionary studies for a single species, or macro-evolutionary in multiple. Since transcriptomes correspond to very functional regions in the genome, there is nucleotide variability under huge selective pressure. Therefore, RNA-Seq is a very information-dense reduced representation of the genome, and valid to study either the effects of micro- and macro-evolution. To prove it, I developed computational pipelines and applied them to discern the population structure of tench, or to obtain the patterns of positive selection in the Clupeiformes order. In both cases, no reference genome assembly was necessary.

5.1 Tench

The study presented in "A novel transcriptome-derived SNPs array for tench (Tinca tinca L.)" (Kumar et al., 2019) is the first one generating genomic and transcriptomic resources for tench. RNA libraries were generated including individuals of both sexes under different metabolic rates, and two different tissues: brain and muscle, in order to maximize transcriptome representation. Genome-wise, ten individuals from six different locations scattered throughout the Eurasian continent were used to maximize genotype variability.

RNA-Seq reads were successfully assembled into a transcriptome, whose quality was assessed through the backmapping of the reads (96.54% to 99.38% success rate) and the search of Actinopterygii SCOs (85.9% retrieved). Additionally, the transcriptome was annotated to obtain the identities of the genes found along with their functions.

Among the 60,414 SNPs identified, 96 were used to design a chip aimed to perform a population genetic study on Tench. This design resulted in a 96% conversion rate, the highest reported to date.

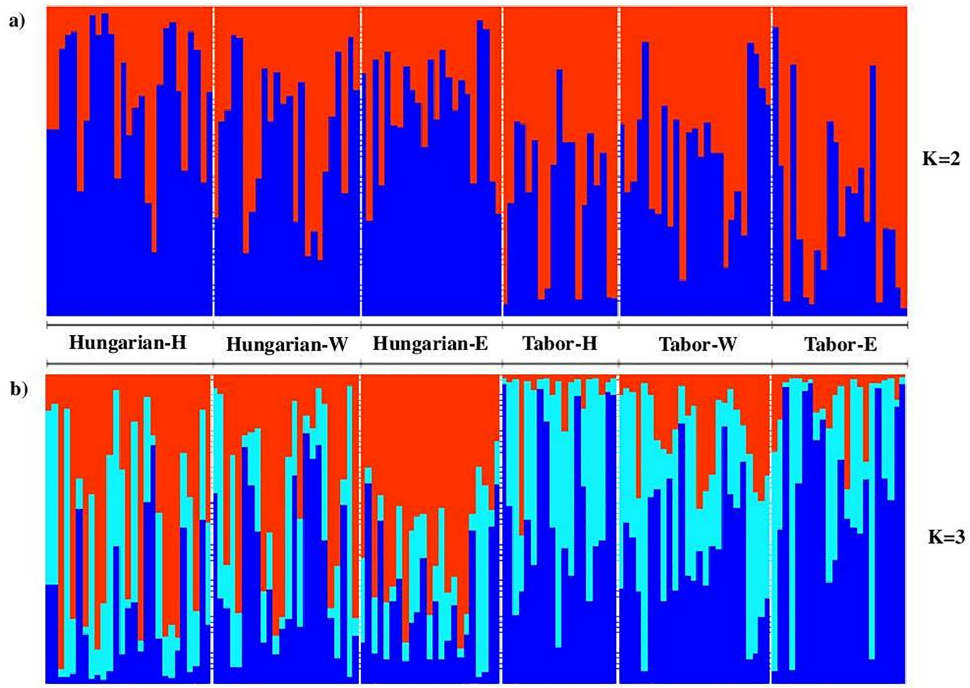


FIGURE 5.1: Structure analysis for tench]. Results from Structure analysis for K = 2 (a) and K = 3 (b). Individuals corresponding to each breed (Hungarian, Tabor) and *gh* gene phylogroup genotype (H: Hybrid; W: Western; E: Eastern) are separated with vertical white bars. These clusters clearly indicate the differences between the two breeds, but not between the *gh* gene genotypes. Figure reproduced from Kumar et al. (2019).

A total of 140 tench fishes, from two different breeds and phylogroup origins, were genotyped with the SNP markers obtained above. On the one hand, clustering results indicate that the most probable number of ancestries is two, which coincides with the number of known phylogroups in this species, Western and Eastern phylogroups (Figure 5.1). Moreover, most individuals showed a mixed ancestry. According to our results, within both Tabor and Hungarian breeds, there were no genetic differences between individuals with Western (W), Eastern (E) or Hybrid (H) genotypes for the gh gene (Table 5.1). This result indicates that within each breed there is gene flow between individuals of both phylogroups, Western and Eastern, supporting the hypothesis that there is no reproductive isolation between the two phylogroups (Kumar et al., 2014), at least in captivity. Beyond this hypothesis, our results would support that, after several generations of mating, individuals of each breed will consist of a mosaic of both phylogroups.

On the other hand, slight differences emerge when comparing individuals of the Tabor and Hungarian breeds. However, six SNPs showed extremely high allelic differences between breeds and were classified as outliers, i.e., under diversifying selection. Among them, two SNPs were located in growth related genes. Knowing that the Western and Eastern phylogroups show a 0.8% divergence in the *gh* (growth

TABLE 5.1: Pairwise F_{ST} (below diagonal) and p-values (above diagonal) among tench breeds (Hungarian, Tabor) and *gh* gene phylogroup genotype (H: Hybrid; W: Western; E: Eastern). F_{ST} value between the two breeds was low but significant. Table reproduced from Kumar et al. (2019).

	Hungarian-H	**Hungarian-W**	**Hungarian-E**	**Tabor-H**	**Tabor-W**	**Tabor-E**
Hungarian-H	-	0.2022	0.1592	0.0000	0.0000	0.0000
Hungarian-W	0.0012	-	0.0429	0.0379	0.0000	0.0000
Hungarian-E	0.0025	0.0083	-	0.0787	0.0504	0.0000
Tabor-H	0.0619*	0.0000	0.0000	-	0.1973	0.3936
Tabor-W	0.0399*	0.0274*	0.0000	0.0054	-	0.0049
Tabor-E	0.0579*	0.0318*	0.0687*	0.0015	0.0218	-

hormone) gene sequence (Kocour & Kohlmann, 2011), the high allelic differentiation observed between the Tabor and Hungarian breeds in growth-related genes led us to hypothesize that the adaptive differences between the two breeds would arise from a differential phylogroup composition at their foundation.

5.2 EXFI

In "EXFI: Exon and splice graph prediction without a reference genome" (Langa et al., 2020), we show the development of EXFI (initials for Exon Finder), a tool to efficiently split a transcriptome into exons, using WGS reads rather than a genome assembly. EXFI relies on Bloom filters, a probabilistic data structure to store and filter k-mers in a fast and memory-efficient way.

We measured the different parameters that affect the accuracy of the procedure: 1) the inclusion of a pre-filtering step, 2) the allowed memory footprint, 3) the k-mer length, and 4) WGS coverage. The results show that first, the pre-filter step decreases the processing step at the same time that improves accuracy (Figure 5.2a). Second,the test over the allowed memory consumption showed that for a 1 Gb genome, 4 GB are more than enough, showing minimal differences to higher quantities, and therefore allowing processing on standard desktops and laptops. Third, the optimal k-mer length was observed to be in the range of 23 to 35 bp, the first focused in Recall, and the second in Precision (Figure 5.2b). Finally, a WGS coverage between 25 and 40x is optimal, and that going beyond only increases both the runtime and the number of errors introduced (Figure 5.2c).

EXFI was compared to ChopStitch, a similar tool, and GMAP, a reference-based transcript aligned, over a wide range of species and libraries. Human and zebrafish simulations were generated to guide the optimal parameter discovery. Figure 5.2d shows the results over the zebrafish dataset. Additionally, the method was tested on Atlantic salmon and Atlantic herring, where WGS samples were pooled, and therefore variants should interfere with exon prediction. Finally, we tested the methods

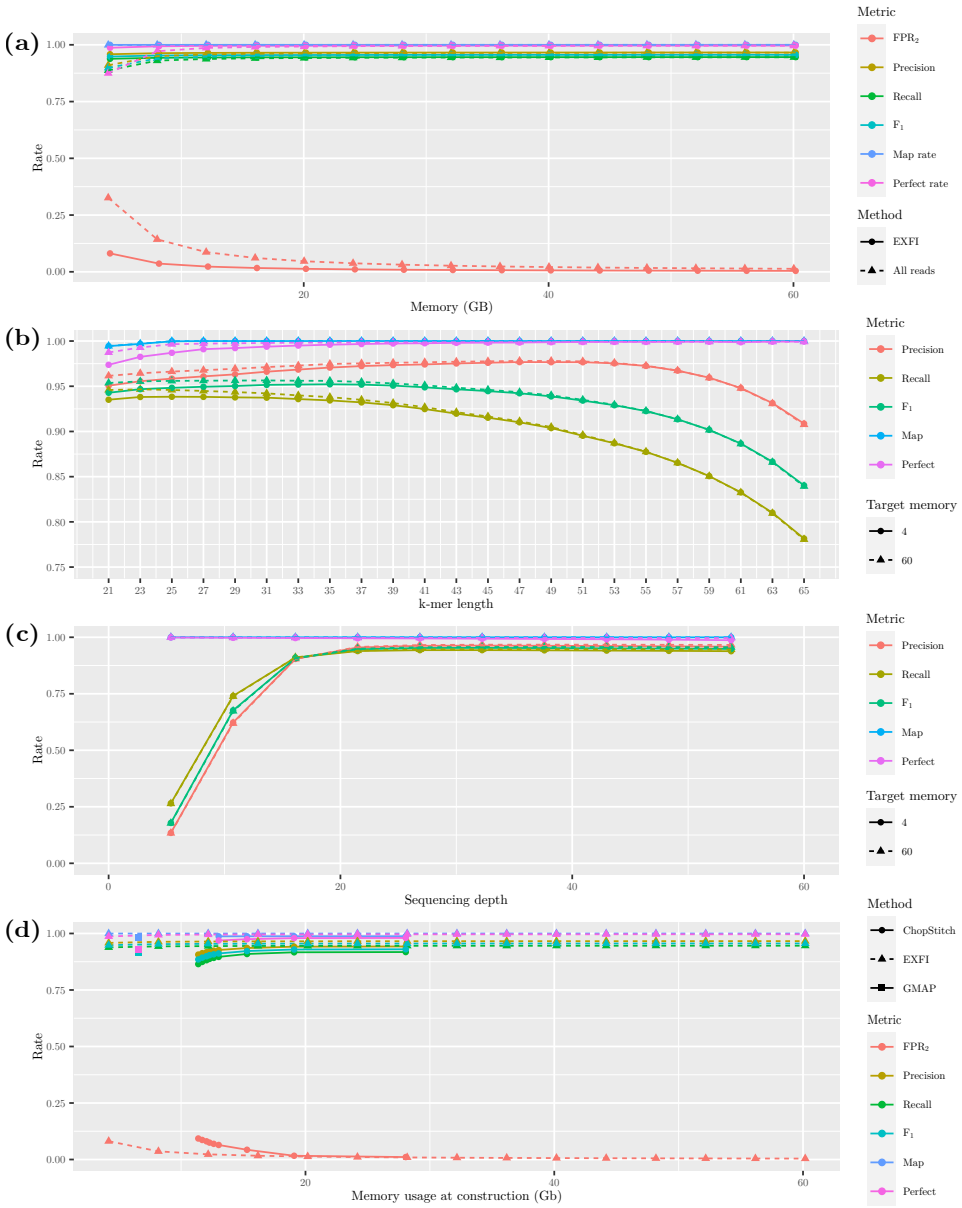


FIGURE 5.2: EXFI results. (a) Filtering and memory. Filtering the WGS resulted in a huge decrease of the BF FPR and therefore higher Precision, Recall and mapping rates. After filtering, memory usage is not an issue anymore. (b) k-mer length. A k-mer length in a window between 25 to 35 resulted in an optimal equilibrium between Precision and Recall. (c) Depth. An optimal depth between 20x to 30x is optimal. After that the method underperforms. (d) Comparison with other tools under the zebrafish dataset. EXFI outperforms in any memory configuration the results from ChopStitch and GMAP.

on the mega-genomes of axolotl and sugar pine (32 and 21 Gbp, respectively), as worst-case scenarios. Overall, GMAP was outperformed in every scenario, even though it has at hand the true genomic sequence. In terms of BF FPR, precision, recall and memory footprint, EXFI outperformed the methods, although ChopStitch was faster (Table 5.2). The reasoning behind the differences in results over the two

methods are the pre-filter step and also the exon processing criteria used to remove false positive bases and exons.

TABLE 5.2: Performance of EXFI and the other two tools across different species. Best metrics across the three methods are marked in bold. Memory, expressed in Gigabytes, represents the peak usage in memory. EXFI obtained the best precision, while ChopStitch obtained better recall. With respect to alignments to the genomes, EXFI obtained the best mapping rates. Extracted from Langa et al. (2020).

Species	Method	Time	Memory	FPR1	FPR2	Precision	Recall	F_1	Mapped	Perfect
Zebrafish	ChopStitch	1h41m54s	28.060	0.010	0.011	0.943	0.918	0.930	0.988	0.980
	EXFI	2h35m15s	**4.177**	0.256	0.081	**0.958**	**0.938**	**0.948**	**1.000**	**0.987**
	GMAP	**40m19**	6.567	—	—	0.918	0.917	0.917	0.982	0.927
Human	ChopStitch	4h28m58s	30.424	0.158	0.100	0.903	0.868	0.886	0.988	**0.969**
	EXFI	6h32m49s	**4.364**	0.361	0.137	**0.931**	**0.893**	**0.912**	**1.000**	0.957
	GMAP	**1h11m25s**	9.301	—	—	0.883	0.884	0.883	0.985	0.907
Salmon	ChopStitch	2h57m38s	8.657	0.010	0.010	0.883	0.887	0.885	0.985	0.975
	EXFI	4h49m37s	**4.466**	0.080	0.042	**0.901**	**0.904**	**0.903**	**0.999**	**0.987**
	GMAP	**1h22m15s**	9.320	—	—	0.809	0.830	0.819	0.979	0.866
Herring ref.	ChopStitch	49m53s	5.679	0.010	0.011	0.819	0.858	0.838	0.974	0.965
	EXFI	1h25m6s	**4.123**	0.064	0.024	0.816	0.866	0.840	**1.000**	**0.995**
	GMAP	**19m8s**	4.707	—	—	**0.949**	**0.941**	**0.945**	0.983	0.935
Herring ass.	ChopStitch	50m2s	5.705	0.010	0.011	—	—	—	0.972	**0.871**
	EXFI	1h32m8s	**4.111**	0.068	0.026	—	—	—	**0.986**	0.823
	GMAP	**37m20s**	6.564	—	—	—	—	—	0.921	0.578
Sugar pine	ChopStitch	—	—	—	—	—	—	—	—	—
	EXFI	2d7h38m57s	60.090	0.090	0.031	—	—	—	**0.997**	**0.903**
	GMAP	**6h20m13s**	**55.371**	—	—	—	—	—	0.956	0.673
Axolotl	ChopStitch	**14h29m38s**	**29.629**	0.202	0.142	—	—	—	0.851	0.772
	EXFI	1d3h20m50s	60.313	0.040	0.020	—	—	—	**0.988**	**0.782**
	GMAP	—	—	—	—	—	—	—	—	—

We applied the method to predict IEB in *T. tinca*, as we did for Objective 1 From the original 266,578 transcripts, 1,072,772 exons were predicted, and from them 228,931 SNPs and 26,169 indels were predicted safe for genotyping. All IEBs proximal to the 96 SNPs used for genotyping were found, showing a 100% Precision, although a single SNP failed to be recognized, indicating also a 98.95% Recall.

In conclusion, EXFI predicts the splice graph from a transcriptome and WGS reads without the need of a reference genome. Multiple parameters were studied and optimized: read filtering, memory usage, k-mer length and sequencing depth. It was tested under a wide number of datasets ranging in depth, heterozygosity, genome length and complexity. A revisit on the tench dataset shows a 100% precision and 99% recall when finding SNPs. Also, low computational resources are needed to carry on the decomposition. Finally, I demonstrate that this tool will be useful in population genetic studies since it is capable of discovering hundreds of thousands of safe-to-genotype SNPs, be useful to design a targeted capture assay (with Exome-Seq in mind), and even expression chips, given that the exon composition is what characterizes transcripts.

5.3 European Sardine

The results of the sequencing effort made to build the European sardine transcriptome profile were presented in "Transcriptomic dataset for Sardina pilchardus" (Langa et al., 2021). Anyway, before sending a boat to catch and dissect the samples, it was necessary to decide what tissues were necessary to collect.

A complex computational study was performed to obtain the most number of transcripts expressed, with the constraint of using a single Illumina HiSeq 2000 lane at the time of our study: 180 million paired-end reads of 100 bp in length, or 36 Gbp. Results over zebrafish showed that tissue diversity is much more important than depth, even at the risk of losing tissue-specific and under expressed transcripts. No matter the sampling depth, the mix of multiple tissues always outperformed any single tissue for the same sequencing effort (Figure 5.3).

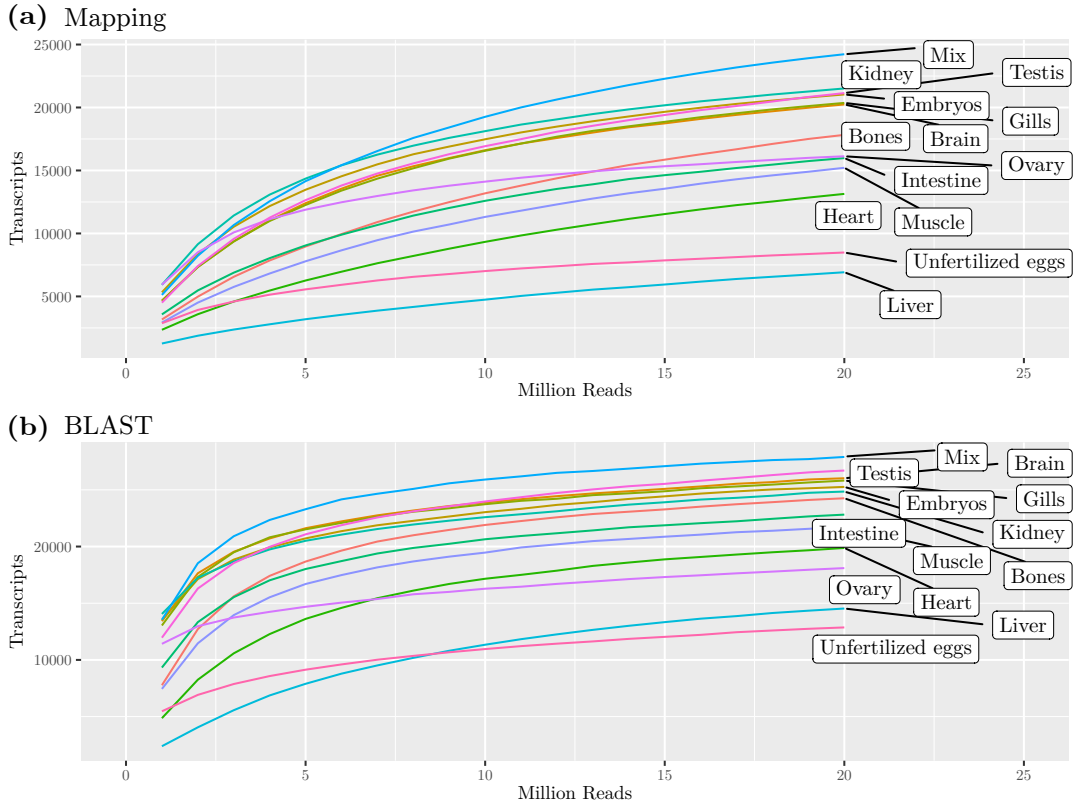


FIGURE 5.3: Retrieved transcripts per library in zebrafish. Reference-guided (a) and *de novo* (b) assembled transcripts are plotted according to the library and sampling used. In both cases, the mix outperforms any tissue for any depth. Also, the rank of every library changes with respect to the methods, but not the mix.

Therefore, we obtained European sardines samples from a scientific bottom trawl survey, from which nine tissues were dissected: brain, eye, heart, kidney, liver, muscle, ovary, skin and testes. The assembly resulted in a transcriptome composed of 198,597 sequences, in which 85% of the Actinopterygii SCOs were found, and that

97.8% of the original reads were placed into it. Annotation resulted in the identification of protein-coding sequences, along with their functions. To our knowledge, this is one of the widest datasets in Clupeiformes but non-model fishs, probably only surpassed the ones from Pasquier et al. (2016), where 13 tissues of 23 fish species were sampled.

This dataset opens the door to studies in population genomics such as the one presented in Objective 1. Also, it will facilitate the study of this species in the aquaculture industry, in areas such as the identification of stocks, the study of traits of interest, and the comparative studies such as the ones done for Objective 4.

5.4 Clupeids

The results here explained will be published in the manuscript titled "Recurrent positive selection of lipid trafficking genes in Clupeiformes" (Langa et al., n.d.). We either obtained or assembled transcriptomes for twelve species of clupeids. The automation investment done in transcriptome assembly and CDS prediction from the previous objectives and publications resulted in a head start in this one. Additional automation in the form of the positive selection pipeline *smsk_selection* (see Section XX) will surely help comparative analysis in any other group of species. In relation with Objective 3, the optimal sampling strategy, transcriptomes derived over a wide number of tissues (*A. pseudoharengus*, *E. encrasicolus* and *S. pilchardus*) performed much better than libraries composed of smaller numbers of tissues but deeply sequenced, validating again the results from Objective 3 (Figure 5.4).

Given the absence of references in Clupeiformes, we had to fall back on a very intensive and conservative transcript clustering into orthogroups, groups of highly similar sequences. This resulted in the grouping of an initial set of one million transcripts into a more manageable set of high-quality 19,914 orthogroups, which correlates to the around 25,000 protein-coding genes present in zebrafish.

Using four-fold degenerate sites, positions where positive selection should be minimal, and a Bayesian framework, we inferred a species tree with a 100% bootstrap support. This tree (Figure 5.5) supports recent studies from Lavoué et al. (2013) and D. Bloom and Egan (2018), based on a few mitochondrial and gene markers, and disproves prior studies based on morphology and parsimony (Nelson et al., 2016).

After obtaining a clear phylogeny of Clupeiformes, we performed another round of orthogroup refinement and an exhaustive search of positive selection in the eight major subdivisions possible in the dataset generated, marked in red in Figure 5.5. In total, we found 918 orthogroups under positive selection. Thanks to the annotation of the orthogroups, we found over-represented six major sections of the genome: 1) the mitochondrial Electron Transport Chain, 2) the ribosomes and the translation

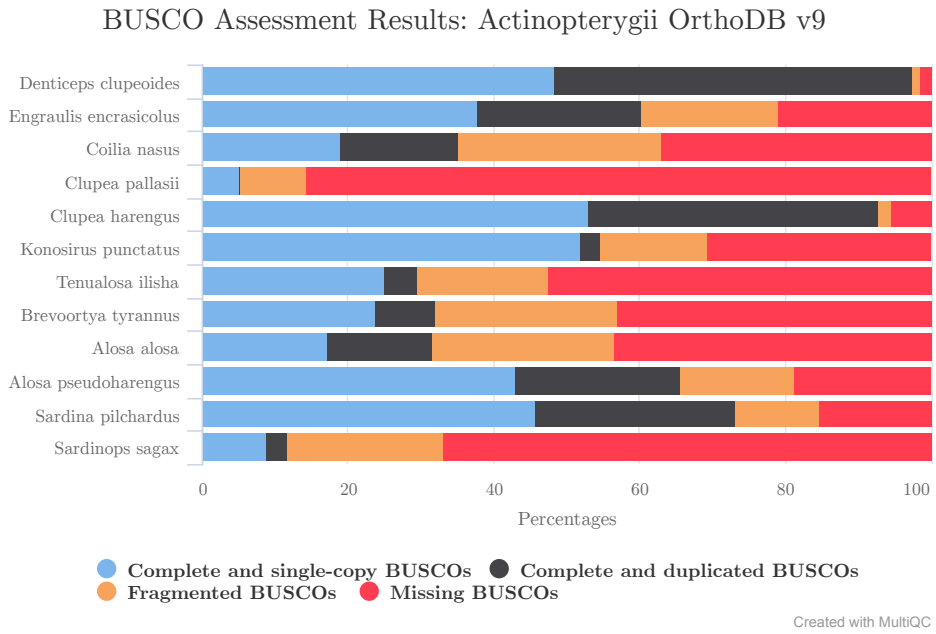


FIGURE 5.4: BUSCO results over Clupeiformes. Transcriptomes derived from assembled genomes obtained the highest results (*D. clupeoides* and *C. harengus*). These were followed by *S. pilchardus*, *A. pseudoharengus*, and *E. encrasicolus*, the ones derived from the widest number of tissues. The set of the Actinopterygii SCOs consists of 4,584 sequences.

machinery, 3) lysosomes, 4) caveolae, 5) cluster of differentiation molecules, and 6) the set of extracellular proteins (Figure 5.6).

Further inspection of the results yielded more important insights. We observed positively selected genes related to the transport and storage of lipids, apolipoproteins, and many of their receptors and mediators, involved in extra- and intracellular cholesterol trafficking. In particular HDL apolipoproteins were found under selection in multiple branches, suggesting selection for the Reverse Cholesterol Transport, the transport of excess cholesterol from peripheral tissues to the liver, where it can be transformed into other cholesterol hormones or removed via the intestines. Additionally, genes were found in the supply of lipids into peripheral tissues via LDLs and intracellular trafficking of these macromolecules. Contrary to what one would expect, most marine species lack the ability to completely synthesize LC-PUFAs (Garrido et al., 2019). However, they present a richness of LC-PUFAs unmatched by any other fish. This paradox, together with the bioenergetic requirements for overwintering, lead us to hypothesize that these species are under enormous selective pressure to store lipids in general, LC-PUFAs in particular, as their evolutionary strategy to store energy and survive. Further studies will be necessary to verify this hypothesis. On a final note, this study has compiled into an automated and user-friendly pipeline the complex methods sections from previous multiple papers. This pipeline is species-agnostic and therefore can be applicable to any other

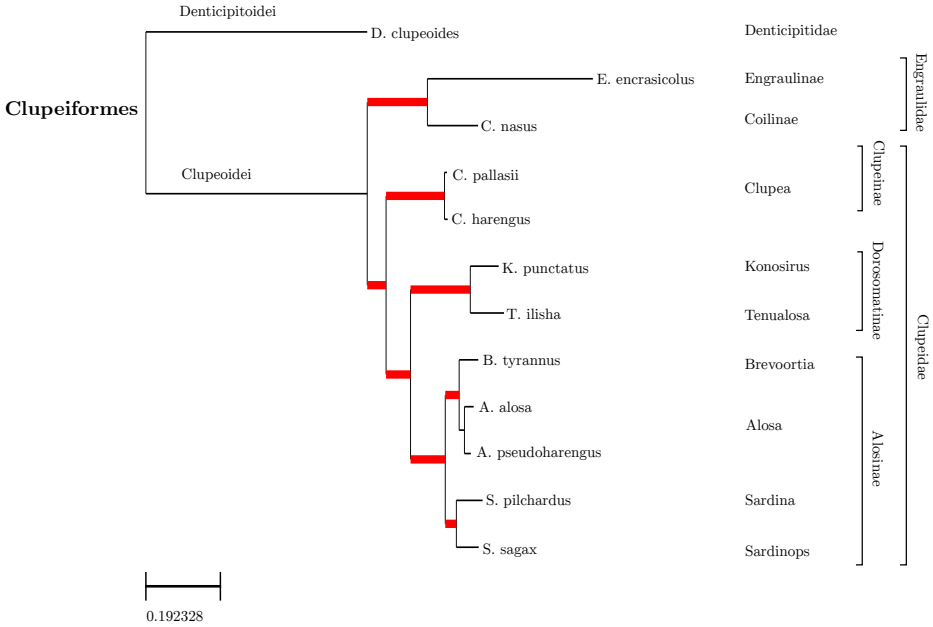


FIGURE 5.5: Bayesian phylogenetic tree of Clupeiformes. Consensus phylogeny inferred by ExaBayes and four-fold degenerate sites. Branch length represents the number of observed mutations. All branches obtained 100% bootstrap support. Branches marked in red are the ones tested for selection in the branch-site test.

taxonomic rank. The elimination of such a great barrier will surely allow any computational biologist to study her own hypotheses.

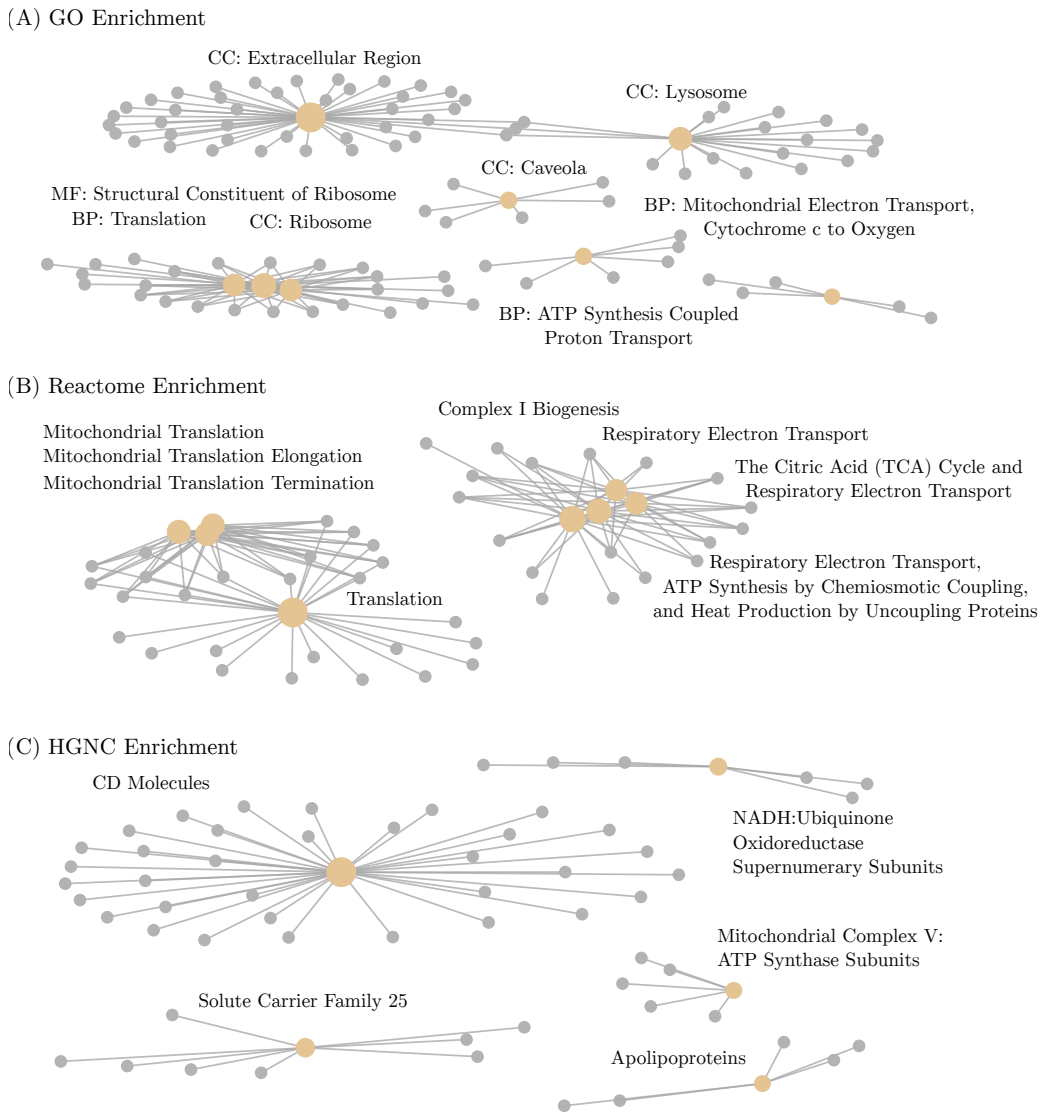


FIGURE 5.6: Overrepresented terms in (a) GO terms, (b) Reactome pathways, (c) HGNC gene families. Overrepresented terms include 1) the Electron Transport Chain, 2) Ribosomes, 3) Lysosomes, 4) Caveolae, 5) CD molecules, and 6) extracellular proteins. Adapted from Langa et al. (n.d.).

5.5 Future Perspectives

Taken together the four publications presented, we can define a multiple stage experimental design to study non-model species.

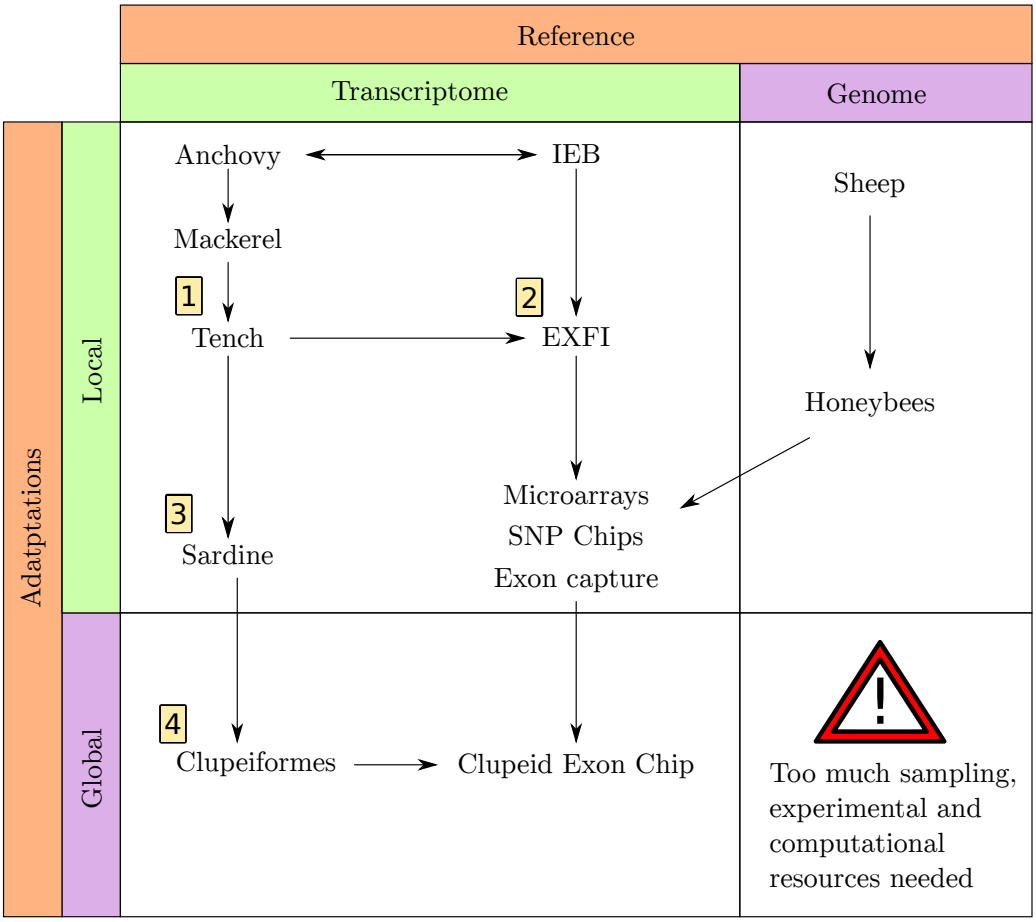


FIGURE 5.7: Schematic representation of this Thesis, along with its possible applications. Parallely to this thesis, we carried out the study of selective sweeps in local breeds of sheep, and to characterize populations of honeybees across the European continent. The use of Pool-Seq is key to account for the total genetic variability. The introduction of EXFI allows us to design not only SNP chips, but to design expression arrays and exon capture chips. Moreover, the application of this approach to multiple species allows the design of exon chips of conserved regions.

In Langa et al. (2021) we demonstrated that a shallow but multi-tissue RNA-Seq study can effectively represent the protein coding part of the genome. Second, EXFI (Langa et al., 2020) effectively decomposes the transcriptome into exons. Therefore, when considering population genetics studies in non-model organisms, in a first stage, we propose to perform shallow RNA-Seq over the widest number of tissues, accompanied by shallow WGS of a few individuals from a small number but diverse

populations. Then, EXFI can be applied to identify polymorphic SNPs with a subsequent genotyping success rate close to 100%. Finally, we can use them to perform population genetics studies, such as the ones done in Kumar et al. (2019). For posterior studies, moving from a micro to a macro scale study, in terms of the number of SNPs, individuals and populations analyzed, two non-exclusive options open in front of us:

1. Design a SNP chip, taking into account all the exons and SNPs discovered, to genotype hundreds to thousands of individuals and/or loci in a single assay.
2. Design an Exome-Seq chip to capture exon fragments and discard intergenic and intronic regions of the genome, and then sequence hundreds of individuals, either individually or in a Pool-Seq fashion, over a much larger number of populations, and then perform Population Genetic studies.

Either approaches appear as alternatives to GBS and RAD-Seq, which rely on restriction enzymes and fragment size selection in order to sequence a small fraction of the genome, and then extract markers from polymorphic markers. The works presented in this thesis, along with the ones that this is based on (Genomic Resources Development Consortium et al., 2015; Montes et al., 2013; see Figure 5.7), they focus on SNPs derived from protein-coding sequences, which are more informative, because they not only report genomic variation, as is the case with most RAD-Seq and GBS sequences, but also on their function. In addition, taken together that most mutations are neutral, and that the protein-coding sequences of the genome is under very selective pressure, the observed SNPs are of great importance in evolutionary studies. Going even further, the use of RNA-Seq and transcriptomes across multiple species can be combined as done in the study of Clupeiformes, whereas the same can not be said for RAD-Seq and GBS since the studied genome-fragments vary from species to species.

Finally, following the approach developed and applied in this Thesis, exon decomposition of multiple species can be achieved over independent studies, and then, they can be integrated to compose an order-specific (or any other taxonomic rank) chip of conserved exons. The purpose of this chip is to sequence both the conserved exons and their flanking regions to apply it to an even higher number of species in a taxon, as done in the Ultra Conserved Elements approach (Faircloth et al., 2012). Following this train of thought, a recent advance in nanopore sequencing (Payne et al., 2021), promises the selective sequencing of regions of the genome. Briefly, nanopores accept or reject the molecule it is sequencing according to a target list. This way, fragments whose start coincide with a known exon are processed, and those who do not are rejected on the spot. Moreover, once a certain coverage for our sequences of interest is achieved, it can start rejecting them too. This technological advance undoubtedly highlights even more the importance of approaches such as EXFI, that provide necessary reference sequences without an initial huge investment.

Another example of our efforts to reduce both economic and computational costs are two studies by our group that used Pool-Seq approaches instead of whole genome individual sequencing. They resulted in two publications made in parallel to this thesis. The first one was aimed to discover selective sweeps in two native sheep breeds of the Basque Country, the Latxa and Sasi Ardi, using classical population genetics statistics (Ruiz-Larrañaga et al., 2018). The latter, in turn, motivated the same Pool-Seq approach to discern honey bee subspecies across Europe (Momeni et al., 2021) in this case using machine learning approaches. Moreover, a third article (Chen et al., n.d.) underscores the power of Pool-Seq versus individual whole genome sequencing: Pool-Seq is much cheaper and the overall genetic population structure is still retrieved.

In summary, this thesis has presented a multidisciplinary effort to study fish species under the light of evolution at two different time-scales. First, I studied the recent adaptations of two cultured breeds of tench. Then, I studied the two immediate improvements to the methodology used: a faster and accurate exon decomposition with EXFI, and a better transcript sampling strategy in the European sardine. Finally, through the combination of multiple datasets, I studied the genes and processes under positive selection in Clupeiformes. Thus, the bioinformatic resources generated in this Thesis together with the results and new hypothesis generated in both evolutionary studies will surely guide future studies, not only in the fishes here studied but also in any other species.

Chapter 6

Bibliography

Aberer, A. J., Kobert, K., & Stamatakis, A. (2014). ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Molecular Biology and Evolution*, *31*(10), 2553–2556. https://doi.org/10.1093/molbev/msu236

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., … Venter, J. C. (2000). The Genome Sequence of *Drosophila melanogaster*. *Science*, *287*(5461), 2185–2195. https://doi.org/10.1126/science.287.5461.2185

Albaina, A., Aguirre, M., Abad, D., Santos, M., & Estonba, A. (2016). 18S rRNA V9 metabarcoding for diet characterization: A critical evaluation with two sympatric zooplanktivorous fish species. *Ecology and Evolution*, *6*(6), 1809–1824. https://doi.org/10.1002/ece3.1986

Alexandrou, M. A., Swartz, B. A., Matzke, N. J., & Oakley, T. H. (2013). Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Molecular Phylogenetics and Evolution*, *69*(3), 514–523. https://doi.org/10.1016/j.ympev.2013.07.026

Altshuler, D., Donnelly, P., & The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Barrio, A. M., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman, D., Höppner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma, K., Zhan, X., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife*, *5*, e12081. https://doi.org/10.7554/eLife.12081

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-Resolution Profiling of Histone Methylations

in the Human Genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, *13*(4), 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bloom, B. H. (1970). Space/Time Trade-offs in Hash Coding with Allowable Errors. *Commun. ACM*, *13*(7), 422–426. https://doi.org/10.1145/362686.362692

Bloom, D., & Egan, J. P. (2018). Systematics of Clupeiformes and testing for ecological limits on species richness in a trans-marine/freshwater clade. *Neotropical Ichthyology*, *16*(3). https://doi.org/10.1590/1982-0224-20180095

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

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., & Crawford, G. E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, *132*(2), 311–322. https://doi.org/10.1016/j.cell.2007.12.014

Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., Yates, B., & Bruford, E. (2019). Genenames.org: The HGNC and VGNC resources in 2019. *Nucleic Acids Research*, *47*(D1), D786–D792. https://doi.org/10.1093/nar/gky930

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Brown, J. W., Walker, J. F., & Smith, S. A. (2017). Phyx: Phylogenetic tools for unix. *Bioinformatics*, *33*(12), 1886–1888. https://doi.org/10.1093/bioinformatics/btx063

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T.-H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Peshkin, L., Tabin, C. J., … Whited, J. L. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, *18*(3), 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

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

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Burdge, G. C. (1998). The role of docosahexaenoic acid in brain development and fetal alcohol syndrome. *Biochemical Society Transactions*, *26*(2), 246–251. https://doi.org/10.1042/bst0260246

Calder, P. C. (2010). Omega-3 Fatty Acids and Inflammatory Processes. *Nutrients*, *2*(3), 355–374. https://doi.org/10.3390/nu2030355

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Checkley, D. M., Asch, R. G., & Rykaczewski, R. R. (2017). Climate, Anchovy, and Sardine. *Annual Review of Marine Science*, *9*(1), 469–493. https://doi.org/10.1146/annurev-marine-122414-033819

Chen, C., Parejo, M., Momeni, J., Langa, J., Nielsen, R. O., Shi, W., Vingborg, R., Kryger, P., Bouga, M., Estonba, A., & Meixner, M. D. (n.d.). Empirical comparison of pool and individual whole-genome sequencing to assess population structure and diversity in European honey bees (Apis mellifera L.)

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R., McPherson, J. D., Miner, T. L., Nash, W. E., Nelson, J. O., Nhan, M. N., Pepin, K. H., Pohl, C. S., Ponce, T. C., Schultz, B., Thompson, J., … Members of the Mouse Genome Analysis Group. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. https://doi.org/10.1038/nature01262

Christoffels, A., Koh, E. G. L., Chia, J.-m., Brenner, S., Aparicio, S., & Venkatesh, B. (2004). Fugu Genome Analysis Provides Evidence for a Whole-Genome Duplication Early During the Evolution of Ray-Finned Fishes. *Molecular Biology and Evolution*, *21*(6), 1146–1151. https://doi.org/10.1093/molbev/msh114

Chu, J., Sadeghi, S., Raymond, A., Jackman, S. D., Nip, K. M., Mar, R., Mohamadi, H., Butterfield, Y. S., Robertson, Gordon, A., & Birol, I. (2014). BioBloom tools: Fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics*, *30*(23), 3402–3404. https://doi.org/10.1093/bioinformatics/btu558

Ciezarek, A. G., Dunning, L. T., Jones, C. S., Noble, L. R., Humble, E., Stefanni, S. S., & Savolainen, V. (2016). Substitutions in the Glycogenin-1 Gene Are Associated with the Evolution of Endothermy in Sharks and Tunas. *Genome Biology and Evolution*, *8*(9), 3011–3021. https://doi.org/10.1093/gbe/evw211

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265–270. https://doi.org/10.1038/nnano.2009.12

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Conklin, D., Montes, I., Albaina, A., & Estonba, A. (2013). Improved conversion rates for SNP genotyping of nonmodel organisms., In *IWBBIO 2013*.

Cormode, G., & Muthukrishnan, S. (2005). An improved data stream summary: The count-min sketch and its applications. *Journal of Algorithms*, *55*(1), 58–75. https://doi.org/10.1016/j.jalgor.2003.12.001

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., Fenton, J., Fenzl, T., Fish, J., Garcia-Gutierrez, L., Garland, P., Gluck, J., González, I., Guermond, S., Guo, J., … Brown, C. T. (2015). The khmer software package: Enabling efficient nucleotide sequence analysis. *F1000Research*. https://doi.org/10.12688/f1000research.6924.1

Czesny, S., Epifanio, J., & Michalak, P. (2012). Genetic Divergence between Freshwater and Marine Morphs of Alewife (Alosa pseudoharengus): A 'Next-Generation' Sequencing Analysis. *PLOS ONE*, *7*(3), e31803. https://doi.org/10.1371/journal.pone.0031803

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294. https://doi.org/10.1093/molbev/msz189

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Divya, B. K., Mohindra, V., Singh, R. K., Yadav, P., Masih, P., & Jena, J. K. (2019). Muscle transcriptome resource for growth, lipid metabolism and immune system in Hilsa shad, Tenualosa ilisha. *Genes & Genomics*, *41*(1), 1–15. https://doi.org/10.1007/s13258-018-0732-y

Dlugosch, K. M., Lai, Z., Bonin, A., Hierro, J., & Rieseberg, L. H. (2013). Allele Identification for Transcriptome-Based Population Genomics in the Invasive Plant

Centaurea solstitialis. *G3: Genes, Genomes, Genetics*, *3*(2), 359–367. https://doi.org/10.1534/g3.112.003871

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … UC Irvine. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Comput Biol*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., … Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Emms, D. M., & Kelly, S. (2017). STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution*, *34*(12), 3267–3278. https://doi.org/10.1093/molbev/msx259

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Eyre-Walker, A., & Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. *Nature*, *397*(6717), 344–347. https://doi.org/10.1038/16915

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, *61*(5), 717–726. https://doi.org/10.1093/sysbio/sys004

FAO. (2019). *FAO Yearbook. Fishery and Aquaculture Statistics 2017/FAO annuaire. Statistiques des pêches et de l'aquaculture 2017/FAO anuario. Estadísticas de pesca y acuicultura 2017*. Rome, FAO.

Fisher, R. (1934). *Statistical methods for research workers, 5th ed.* Oliver, Boyd, Edinburgh.

Foll, M., & Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Francis, R. M. (2017). Pophelper : An R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, *17*(1), 27–32. https://doi.org/10.1111/1755-0998.12509

Garrido, D., Kabeya, N., Betancor, M. B., Pérez, J. A., Acosta, N. G., Tocher, D. R., Rodríguez, C., & Monroig, Ó. (2019). Functional diversification of teleost Fads2 fatty acyl desaturases occurs independently of the trophic level. *Scientific Reports*, *9*(1), 11199. https://doi.org/10.1038/s41598-019-47709-0

Genomic Resources Development Consortium, Álvarez, P., Arthofer, W., Coelho, M. M., Conklin, D., Estonba, A., Grosso, A. R., Helyar, S. J., Langa, J., Machado, M. P., Montes, I., Pinho, J., Rief, A., Schartl, M., Schlick-Steiner, B. C., Seeber, J., Steiner, F. M., & Vilas, C. (2015). Genomic Resources Notes Accepted 1 June 2015 – 31 July 2015. *Molecular Ecology Resources*, *15*(6), 1510–1512. https://doi.org/10.1111/1755-0998.12454

Gharib, W. H., & Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Molecular Biology and Evolution*, *30*(7), 1675–1686. https://doi.org/10.1093/molbev/mst062

Glasauer, S. M. K., & Neuhauss, S. C. F. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics*, *289*(6), 1045–1060. https://doi.org/10.1007/s00438-014-0889-2

Gouveia-Oliveira, R., Sackett, P. W., & Pedersen, A. G. (2007). MaxAlign: Maximizing usable data in an alignment. *BMC Bioinformatics*, *8*(1), 312. https://doi.org/10.1186/1471-2105-8-312

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., Palma, F. d., 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

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Haas, B. J. (2016). TransDecoder (Find Coding Regions Within Transcripts). Retrieved June 27, 2016, from https://transdecoder.github.io/

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

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249. https://doi.org/10.1016/S1074-5521(98)90108-9

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (2016). Ensembl comparative genomics resources. *Database*, *2016*(bav096). https://doi.org/10.1093/database/bav096

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J., & McCombie, W. R. (2007). Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, *39*(12), 1522–1527. https://doi.org/10.1038/ng.2007.42

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. https://doi.org/10.1038/nature12111

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering,

C., & Bork, P. (2016). eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293. https://doi.org/10.1093/nar/gkv1248

Huret, M., Lebigre, C., Iriondo, M., Montes, I., & Estonba, A. (2020). Genetic population structure of anchovy (Engraulis encrasicolus) in North-western Europe and variability in the seasonal distribution of the stocks. *Fisheries Research*, *229*, 105619. https://doi.org/10.1016/j.fishres.2020.105619

ICES. (2018). Report of the Working Group on Southern Horse Mackerel, Anchovy and Sardine (WGHANSA), Lisbon, Portugal, ICES. http://www.ices.dk/community/groups/Pages/WGhansa.aspx

í Kongsstovu, S., Mikalsen, S.-O., Homrum, E. í., Jacobsen, J. A., Flicek, P., & Dahl, H. A. (2019). Using long and linked reads to improve an Atlantic herring (Clupea harengus) genome assembly. *Scientific Reports*, *9*(1), 17716. https://doi.org/10.1038/s41598-019-54151-9

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science (New York, N.Y.)*, *324*(5924), 218. https://doi.org/10.1126/science.1168978

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, *21*(7), 1160–1167. https://doi.org/10.1101/gr.110882.110

Iv, F. J. Z., Rana, S. B., Alvi, Z. A., Zhang, Z., Murphy, W., & Bentivegna, C. S. (2017). De Novo Assembly and Analysis of the Testes Transcriptome from the Menhaden, Bervoortia tyrannus. *Fisheries and Aquaculture Journal*, *8*(1), 1–8. https://doi.org/10.4172/2150-3508.1000186

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., & Birol, I. (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777. https://doi.org/10.1101/gr.214346.116

Jain, A. P., Aggarwal, K. K., & Zhang, P.-Y. (2015). Omega-3 fatty acids and cardiovascular disease. *European Review for Medical and Pharmacological Sciences*, *19*(3), 441–445.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, *316*(5830), 1497–1502. https://doi.org/10.1126/science.1141319

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Keinath, M. C., Timoshevskiy, V. A., Timoshevskaya, N. Y., Tsonis, P. A., Voss, S. R., & Smith, J. J. (2015). Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing. *Scientific Reports*, *5*, 16413. https://doi.org/10.1038/srep16413

Khan, H., Mohamadi, H., Vandervalk, B. P., Warren, R. L., Chu, J., & Birol, I. (2018). ChopStitch: Exon annotation and splice graph construction using transcriptome assembly and whole genome sequencing data. *Bioinformatics*, *34*(10), 1697–1704. https://doi.org/10.1093/bioinformatics/btx839

Kijas, J., McWilliam, S., Sanchez, M. N., Kube, P., King, H., Evans, B., Nome, T., Lien, S., & Verbyla, K. (2018). Evolution of Sex Determination Loci in Atlantic Salmon. *Scientific Reports*, *8*(1), 5664. https://doi.org/10.1038/s41598-018-23984-1

Kim, S.-K., & Mendis, E. (2006). Bioactive compounds from marine processing byproducts – A review. *Food Research International*, *39*(4), 383–393. https://doi.org/10.1016/j.foodres.2005.10.010

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., & Flicek, P. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation*, *2011*, bar030. https://doi.org/10.1093/database/bar030

Kocour, M., & Kohlmann, K. (2014). Distribution of five growth hormone gene haplogroups in wild and cultured tench, Tinca tinca L., populations. *Journal of Applied Ichthyology*, *30*(s1), 22–28. https://doi.org/10.1111/jai.12428

Kocour, M., Gela, D., Rodina, M., & Flajšhans, M. (2010). Performance of different tench, Tinca tinca (L.), groups under semi-intensive pond conditions: It is worth establishing a coordinated breeding program. *Reviews in fish biology and fisheries*, *20*(3), 345–355.

Kocour, M., & Kohlmann, K. (2011). Growth hormone gene polymorphisms in tench, Tinca tinca L. *Aquaculture*, *310*(3), 298–304. https://doi.org/10.1016/j.aquaculture.2010.10.006

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Kumar, G., Kohlmann, K., Gela, D., & Kocour, M. (2014). Phylogroup origin of Tench Tinca tinca L. has no effects on main performance parameters., San Sebastián.

Kumar, G., Langa, J., Montes, I., Conklin, D., Kocour, M., Kohlmann, K., & Estonba, A. (2019). A novel transcriptome-derived SNPs array for tench (Tinca tinca L.) *PLOS ONE*, *14*(3), e0213992. https://doi.org/10.1371/journal.pone.0213992

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, *12*(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

Lajbner, Z., Linhart, O., & Kotlík, P. (2011). Human-aided dispersal has altered but not erased the phylogeography of the tench. *Evolutionary Applications*, *4*(4), 545–561. https://doi.org/10.1111/j.1752-4571.2010.00174.x

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundström, G., Rubin, C.-J., Gilbert, E. R., Berglund, J., Wetterbom, A., Laikre, L., Webster, M. T., Grabherr, M., Ryman, N., & Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, *109*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Langa, J., Estonba, A., & Conklin, D. (2020). EXFI: Exon and splice graph prediction without a reference genome. *Ecology and Evolution*, *10*(16), 8880–8893. https://doi.org/https://doi.org/10.1002/ece3.6587

Langa, J., Huret, M., Montes, I., Conklin, D., & Estonba, A. (2021). Transcriptomic dataset for Sardina pilchardus: Assembly, annotation, and expression of nine tissues. *Data in Brief*, 107583. https://doi.org/10.1016/j.dib.2021.107583

Langa, J., Rueda, Y., Albaina, A., Huret, M., Conklin, D., & Estonba, A. (n.d.). Recurrent positive selection of lipid trafficking genes in Clupeiformes, Manuscript under review on Marine Biotechnology.

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

Lassmann, T., Frings, O., & Sonnhammer, E. L. L. (2009). Kalign2: High-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Research*, *37*(3), 858–865. https://doi.org/10.1093/nar/gkn1006

Lavoué, S., Miya, M., Musikasinthorn, P., Chen, W.-J., & Nishida, M. (2013). Mitogenomic Evidence for an Indo-West Pacific Origin of the Clupeoidei (Teleostei: Clupeiformes) (F. Tinti, Ed.). *PLoS ONE*, *8*(2), e56485. https://doi.org/10.1371/journal.pone.0056485

Leaute, J.-P., Pawloski, L., & Salaun, M. (2015). EVHOE 2015 cruise,Thalassa R/V. https://doi.org/10.17600/15002200

Lemaitre, R. N., King, I. B., Mozaffarian, D., Kuller, L. H., Tracy, R. P., & Siscovick, D. S. (2003). N-3 Polyunsaturated fatty acids, fatal ischemic heart disease, and nonfatal myocardial infarction in older adults: The Cardiovascular

Health Study. *The American Journal of Clinical Nutrition*, *77*(2), 319–325. https://doi.org/10.1093/ajcn/77.2.319

Li, H. (2011). Wgsim: Reads simulator. https://github.com/lh3/wgsim

Li, H. (2012). Seqtk: Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*. Retrieved March 16, 2018, from http://arxiv.org/abs/1303.3997

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

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

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

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology (Clifton, N.J.)*, *1079*, 155–170. https://doi.org/10.1007/978-1-62703-646-7_10

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226

Machado, A. M., Tørresen, O. K., Kabeya, N., Couto, A., Petersen, B., Felício, M., Campos, P. F., Fonseca, E., Bandarra, N., Lopes-Marques, M., Ferraz, R., Ruivo, R., Fonseca, M. M., Jentoft, S., Monroig, Ó., Da Fonseca, R. R., & C. Castro, L. F. (2018). "Out of the Can": A Draft Genome Assembly, Liver Transcriptome, and Nutrigenomics of the European Sardine, Sardina pilchardus. *Genes*, *9*(10), 485. https://doi.org/10.3390/genes9100485

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro,

J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Markova-Raina, P., & Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Research*, *21*(6), 863–874. https://doi.org/10.1101/gr.115949.110

McKinney, W. (n.d.). Pandas: A Foundational Python Library for Data Analysis and Statistics, 9.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., & Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, *454*(7205), 766–770. https://doi.org/10.1038/nature07107

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., … Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, *448*(7153), 553–560. https://doi.org/10.1038/nature06008

Momeni, J., Parejo, M., Nielsen, R. O., Langa, J., Montes, I., Papoutsis, L., Farajzadeh, L., Bendixen, C., Căuia, E., Charrière, J.-D., Coffey, M. F., Costa, C., Dall'Olio, R., De la Rúa, P., Drazic, M. M., Filipi, J., Galea, T., Golubovski, M., Gregorc, A., … Estonba, A. (2021). Authoritative subspecies diagnosis tool for European honey bees based on ancestry informative SNPs. *BMC Genomics*, *22*(1), 1–12. https://doi.org/10.1186/s12864-021-07379-7

Montes, I., Conklin, D., Albaina, A., Creer, S., Carvalho, G. R., Santos, M., & Estonba, A. (2013). SNP Discovery in European Anchovy (Engraulis encrasicolus , L) by High-Throughput Transcriptome and Genome Sequencing. *PLOS ONE*, *8*(8), e70051. https://doi.org/10.1371/journal.pone.0070051

Moretti, S., Laurenczy, B., Gharib, W. H., Castella, B., Kuzniar, A., Schabauer, H., Studer, R. A., Valle, M., Salamin, N., Stockinger, H., & Robinson-Rechavi, M. (2014). Selectome update: Quality control and computational improvements to a database of positive selection. *Nucleic Acids Research*, *42*(D1), D917–D921. https://doi.org/10.1093/nar/gkt1065

Morin, P. A., Luikart, G., Wayne, R. K., & the SNP workshop group. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*(4), 208–216. https://doi.org/10.1016/j.tree.2004.01.009

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304. Retrieved September 28, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461236/

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., Cardeno, C., Koriabine, M., Holtz-Morris, A. E., Liechty, J. D., Martínez-García, P. J., Vasquez-Gross, H. A., Lin, B. Y., Zieve, J. J., Dougherty, W. M., Fuentes-Soriano, S., Wu, L.-S., Gilbert, D., Marçais, G., … Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*(3), R59. https://doi.org/10.1186/gb-2014-15-3-r59

Nelson, J. S., Grande, T. C., & Wilson, M. V. H. (2016). *Fishes of the World: Nelson/Fishes of the World*. Hoboken, NJ, USA, John Wiley & Sons, Inc. https://doi.org/10.1002/9781119174844

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

Pasquier, J., Cabau, C., Nguyen, T., Jouanno, E., Severac, D., Braasch, I., Journot, L., Pontarotti, P., Klopp, C., Postlethwait, J. H., Guiguen, Y., & Bobe, J. (2016). Gene evolution and gene expression after whole genome duplication in fish: The PhyloFish database. *BMC Genomics*, *17*(1), 368. https://doi.org/10.1186/s12864-016-2709-z

Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B. J., & Loose, M. (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology*, *39*(4), 442–450. https://doi.org/10.1038/s41587-020-00746-x

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Petit, E., Balloux, F., & Goudet, J. (2001). Sex-Biased Dispersal in a Migratory Bat: A Characterization Using Sex-Specific Demographic Parameters. *Evolution*, *55*(3), 635–640. https://doi.org/10.1111/j.0014-3820.2001.tb00797.x

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, *14*(7), 687–690. https://doi.org/10.1038/nmeth.4324

Piry, S., Alapetite, A., Cornuet, J.-M., Paetkau, D., Baudouin, L., & Estoup, A. (2004). GENECLASS2: A Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity*, *95*(6), 536–539. https://doi.org/10.1093/jhered/esh074

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Proux, E., Studer, R. A., Moretti, S., & Robinson-Rechavi, M. (2009). Selectome: A database of positive selection. *Nucleic Acids Research*, *37*(suppl_1), D404–D407. https://doi.org/10.1093/nar/gkn768

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

Quinlan, A. R., Stewart, D. A., Strömberg, M. P., & Marth, G. T. (2008). Pyrobayes: An improved base caller for SNP discovery in pyrosequences. *Nature Methods*, *5*(2), 179–181. https://doi.org/10.1038/nmeth.1172

R Core Team. (2020). *R: A language and environment for statistical computing* (manual). Vienna, Austria. https://www.R-project.org/

Reback, J., Jbrockmendel, McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Hawkins, S., Gfyoung, Roeschke, M., Sinhrks, Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Hoefler, P., Naveh, S., Garcia, M., Schendel, J., … Seabold, S. (2021). Pandas-dev/pandas: Pandas 1.3.3. Zenodo. https://doi.org/10.5281/ZENODO.3509134

Redelings, B. (2014). Erasing Errors due to Alignment Ambiguity When Estimating Positive Selection. *Molecular Biology and Evolution*, *31*(8), 1979–1993. https://doi.org/10.1093/molbev/msu174

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rice, W. R. (1989). Analyzing Tables of Statistical Tests. *Evolution*, *43*(1), 223–225. https://doi.org/10.2307/2409177

Richards, D. J., Renaud, L., Agarwal, N., Starr Hazard, E., Hyde, J., & Hardiman, G. (2018). De Novo Hepatic Transcriptome Assembly and Systems Level Analysis of Three Species of Dietary Fish, Sardinops sagax, Scomber japonicus, and Pleuronichthys verticalis. *Genes*, *9*(11), 521. https://doi.org/10.3390/genes9110521

Roberts, S. B., Hauser, L., Seeb, L. W., & Seeb, J. E. (2012). Development of Genomic Resources for Pacific Herring through Targeted Transcriptome Pyrosequencing. *PLOS ONE*, *7*(2), e30908. https://doi.org/10.1371/journal.pone.0030908

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., & Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, *4*(8), 651–657. https://doi.org/10.1038/nmeth1068

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R.,

... Birol, I. (2010). *De novo* assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–912. https://doi.org/10.1038/nmeth.1517

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, *12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

Rousset, F. (2008). Genepop'007: A complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*(1), 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Roux, J., Privman, E., Moretti, S., Daub, J. T., Robinson-Rechavi, M., & Keller, L. (2014). Patterns of Positive Selection in Seven Ant Genomes. *Molecular Biology and Evolution*, *31*(7), 1661–1685. https://doi.org/10.1093/molbev/msu141

Ruiz-Larrañaga, O., Langa, J., Rendo, F., Manzano, C., Iriondo, M., & Estonba, A. (2018). Genomic selection signatures in sheep from the Western Pyrenees. *Genetics Selection Evolution*, *50*, 9. https://doi.org/10.1186/s12711-018-0378-x

Ruxton, C. H. S., Reed, S. C., Simpson, M. J. A., & Millington, K. J. (2004). The health benefits of omega-3 polyunsaturated fatty acids: A review of the evidence. *Journal of Human Nutrition and Dietetics*, *17*(5), 449–459. https://doi.org/10.1111/j.1365-277X.2004.00552.x

Salikhov, K., Sacomoto, G., & Kucherov, G. (2014). Using cascading Bloom filters to improve the memory usage for de Brujin graphs. *Algorithms for Molecular Biology*, *9*(1), 2. https://doi.org/10.1186/1748-7188-9-2

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015). GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, *43*(W1), W7–W14. https://doi.org/10.1093/nar/gkv318

Sidhu, K. S. (2003). Health benefits and potential risks related to consumption of fish or fish oil. *Regulatory Toxicology and Pharmacology*, *38*(3), 336–344. https://doi.org/10.1016/j.yrtph.2003.07.002

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*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815. https://doi.org/10.1038/35048692

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. *Nature*, *449*(7164), 804–810. https://doi.org/10.1038/nature06244

Uauy, R., & Valenzuela, A. (2000). Marine oils: The health benefits of n-3 fatty acids. *Nutrition*, *16*(7), 680–684. https://doi.org/10.1016/S0899-9007(00)00326-9

Valle, M., Schabauer, H., Pacher, C., Stockinger, H., Stamatakis, A., Robinson-Rechavi, M., & Salamin, N. (2014). Optimization strategies for fast detection of positive selection on phylogenetic trees. *Bioinformatics*, *30*(8), 1129–1137. https://doi.org/10.1093/bioinformatics/btt760

Vandepoele, K., Vos, W. D., Taylor, J. S., Meyer, A., & Peer, Y. V. d. (2004). Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences*, *101*(6), 1638–1643. https://doi.org/10.1073/pnas.0307968100

van Dongen, S., & Abreu-Goodger, C. (2012). Using MCL to Extract Clusters from Networks. In J. van Helden, A. Toussaint, & D. Thieffry (Eds.), *Bacterial Molecular Networks: Methods and Protocols* (pp. 281–295). New York, NY, Springer. https://doi.org/10.1007/978-1-61779-361-5_15

Wallace, I. M., O'Sullivan, O., Higgins, D. G., & Notredame, C. (2006). M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *34*(6), 1692–1699. https://doi.org/10.1093/nar/gkl091

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Welcomme, R. L. (1988). *International introductions of inland aquatic species*. Rome, Food; Agriculture Organization of the United Nations.

Whelan, S., & Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics (Oxford, England)*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'* (manual). https://CRAN.R-project.org/package=tidyverse

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

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.-S., & Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Yang, Y., & Smith, S. A. (2014). Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution*, *31*(11), 3081–3092. https://doi.org/10.1093/molbev/msu245

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., & dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, *28*(3), 1217–1228. https://doi.org/10.1093/molbev/msq303

Yokoyama, M., Origasa, H., Matsuzaki, M., Matsuzawa, Y., Saito, Y., Ishikawa, Y., Oikawa, S., Sasaki, J., Hishida, H., Itakura, H., Kita, T., Kitabatake, A., Nakaya, N., Sakata, T., Shimada, K., & Shirato, K. (2007). Effects of eicosapentaenoic acid on major coronary events in hypercholesterolaemic patients (JELIS): A randomised open-label, blinded endpoint analysis. *The Lancet*, *369*(9567), 1090–1098. https://doi.org/10.1016/S0140-6736(07)60527-3

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zarraonaindia, I., Iriondo, M., Albaina, A., Pardo, M. A., Manzano, C., Grant, W. S., Irigoien, X., & Estonba, A. (2012). Multiple SNP Markers Reveal Fine-Scale Population and Deep Phylogeographic Structure in European Anchovy (Engraulis encrasicolus L.) *PLOS ONE*, *7*(7), e42201. https://doi.org/10.1371/journal.pone.0042201

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution*, *22*(12), 2472–2479. https://doi.org/10.1093/molbev/msi237

Zhu, G., Wang, L., Tang, W., Wang, X., & Wang, C. (2017). Identification of olfactory receptor genes in the Japanese grenadier anchovy Coilia nasus. *Genes & Genomics*, *39*(5), 521–532. https://doi.org/10.1007/s13258-017-0517-8

Part II

Conclusions

Tench

- Based essentially on the genetic homogeneity observed between individuals with Western, Eastern and Hybrid *gh* genotypes within breeds we hypothesise that *Tinca tinca* cultured individuals in Central Europe would be mosaics of the Western and Eastern phylogroups, while adaptive differences observed when comparing the Hungarian and Tabor breeds would support a differential phylogroup composition at their foundation

EXFI

- The EXFI computational tool developed in this Thesis partitions a transcriptome into exons while avoiding genome assembly. It uses probabilistic data structures, and it performs the best in terms of memory footprint, accuracy, retrieval and exon mapping in almost all scenarios and metrics, regardless of ploidy or genome size, and shows an accuracy of a perfect mapping of >98% in fish species, making it accurate enough to design exome-level SNP chips.

Sardine

- To obtain a comprehensive transcriptome in a non-model species, we recommend using the maximum number of tissues and the minimum number of individuals. Our results make it clear that, regardless of sampling depth, multiple tissues always outperformed any single tissue for the same sequencing effort.
- This Thesis provides one of the most complete transcriptomes in non-model species: an assembled and annotated version of the European sardine. It also provides the expression profiles of nine tissues. These genomic resources have great applicability in studies aimed at investigating the adaptive mechanisms and sustainable management of this species with such an important ecological and economical impact.

Clupeids

- In this study, a dataset of 12 Clupeiformes transcriptomes was generated to perform the comparative analysis that successfully resolved the phylogeny of the Clupeiformes supporting recently published phylogenetic trees based on a few nuclear and mitochondrial genes, and at the same time disagreeing with previous studies based on morphological characters.
- The phylogenetic tree constructed in this Thesis also served to identify 918 genes under positive selection, which overrepresented six major sections of the genome: the mitochondrial electron transport chain, ribosomes, lysosomes, caveolae, CD molecules and extracellular proteins.
- Positive selection was observed especially focused on genes related to extra- and intracellular lipid trafficking. This is particularly interesting because of the high lipid content in Clupeiformes and their high bioenergetic requirements during hibernation. All this suggests that in Clupeiformes lipid storage is an evolutionary driver for energy storage to ensure survival.

General Conclusions

- This PhD shows through two case studies that RNA-Seq, combined with fast and accurate exon decomposition and extensive gene sampling, provides an efficient surface on which to base phylogeographic and evolutionary studies of non-model organisms. In addition, being a cheap and efficient approach, resources can be reallocated by investing them in the number of populations and species studied.
- Based on the proposed multi-species RNA-Seq strategy, it is possible to design targeted capture chips from conserved exons, which would facilitate, cheapen and increase the resolution of microevolutionary and macroevolutionary studies in non- model species.

Part III

Appendixes

Appendix A

Article 1. Tench

Kumar, G., Langa, J., Montes, I., Conklin, D., Kocour, M., Kohlmann, K., & Estonba, A. (2019). A novel transcriptome-derived SNPs array for tench (Tinca tinca L.). *PLOS ONE*, *14*(3), e0213992. https://doi.org/10.1371/journal.pone.0213992.

Article co-authored with Dr. Girish Kumar.

TABLE A.1: Quality Metrics for PLoS ONE in 2019.

PLoS ONE 2019 - Web of Science	
Category	Multidisciplinary Sciences
Impact Factor	2.740
Rank	27/71
Quantile	Q2

PLoS ONE 2019 - Scopus	
Category	Multidisciplinary
CiteScore	5.2
Rank	10/111
Quantile	Q1

PLOS ONE

RESEARCH ARTICLE

A novel transcriptome-derived SNPs array for tench (*Tinca tinca* L.)

Girish Kumar[1☯*], Jorge Langa[2☯], Iratxe Montes[2], Darrell Conklin[3,4], Martin Kocour[1], Klaus Kohlmann[5], Andone Estonba[2]

1 Research Institute of Fish Culture and Hydrobiology, South Bohemian Research Center of Aquaculture and Biodiversity of Hydrocenoses, Faculty of Fisheries and Protection of Waters, University of South Bohemia in Ceske Budejovice, Czech Republic, **2** Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country UPV/EHU, Leioa-Bilbao, Bizkaia, Spain, **3** Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Gipuzkoa, Spain, **4** IKERBASQUE, Basque Foundation for Science, Bilbao, Bizkaia, Spain, **5** Department of Aquaculture and Ecophysiology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

☯ These authors contributed equally to this work.
* girishkumar.nio@gmail.com

OPEN ACCESS

Citation: Kumar G, Langa J, Montes I, Conklin D, Kocour M, Kohlmann K, et al. (2019) A novel transcriptome-derived SNPs array for tench (*Tinca tinca* L.). PLoS ONE 14(3): e0213992. https://doi.org/10.1371/journal.pone.0213992

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: October 2, 2018

Accepted: March 5, 2019

Published: March 19, 2019

Data Availability Statement: Transcriptome has been uploaded to NCBI Transcriptome Shotgun Assembly Sequence Database and it is available at GenBank with accession number GFZX00000000.1. Polymorphic SNPs have been uploaded to EBI's European Variation Archive under the study accession number PRJEB23783.

Funding: This research was supported by projects CENAKVA and Reproductive and genetic approaches for fish biodiversity conservation and aquaculture (CZ02.1.01/0.0/0.0/16_025/0007370) funded by Ministry of Education, Youth and Sports

Abstract

Tench (*Tinca tinca* L.) has great economic potential due to its high rate of fecundity and long-life span. Population genetic studies based on allozymes, microsatellites, PCR-RFLP and sequence analysis of genes and DNA fragments have revealed the presence of Eastern and Western phylogroups. However, the lack of genomic resources for this species has complicated the development of genetic markers. In this study, the tench transcriptome and genome were sequenced by high-throughput sequencing. A total of 60,414 putative SNPs were identified in the tench transcriptome using a computational pipeline. A set of 96 SNPs was selected for validation and a total of 92 SNPs was validated, resulting in the highest conversion and validation rate for a non-model species obtained to date (95.83%). The validated SNPs were used to genotype 140 individuals belonging to two tench breeds (Tabor and Hungarian), showing low (F_{ST} = 0.0450) but significant (<0.0001) genetic differentiation between the two tench breeds. This implies that set of validated SNPs array can be used to distinguish the tench breeds and that it might be useful for studying a range of associations between DNA sequence and traits of importance. These genomic resources created for the tench will provide insight into population genetics, conservation fish stock management, and aquaculture.

Introduction

Tench (*Tinca tinca* L.) is a freshwater fish species within the *Cyprinidae* family that spawns and grows ideally at water temperatures of 20–29˚C [1, 2]. Its native distribution is Eurasia; however, due to human-mediated movement, tench can also be found in temperate and tropic freshwater regions across the globe [3]. Due to its attractive appearance and specific meat flavour, tench has relevant economic importance and is commonly used in aquaculture and

of the Czech Republic, and by the Genomic Resources Research Group from the Basque University System (IT558-10) funded by the Department of Education, Universities and Research of the Basque Government. JL is supported by the pre-doctoral program Education Department of the Basque Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

sport fishing [4]. For example, tench farming is a common aquaculture activity in Europe and has recently expanded to China [5]. All of these facts motivate the increase of its annual global aquaculture production [6] of about 1400 tons [7].

Tench has very interesting features that set the species apart from other members of the *Cyprinidae* and that have popularized tench as an experimental model [8]. These include: an unequivocal body colour, normally green to brown-green, with golden, blue and albinotic phenotypes also existing [9]; small and hardly visible scales deeply embedded into the dermis; obvious sexual dimorphism in pelvic fins [4], specific reproductive biology [1]; low incidence of viral and bacterial diseases but high susceptibility to some chemical compounds [10]; and monophyletic origin (all descendants of a common ancestor) within *Tinca* genus [11]. Genetics studies have also shown that tench is still a diploid species (2n = 48) [12], which is advantageous for some genetic studies, compared to many cyprinids that are polyploid species [13].

Genetic studies on tench have until now been based on allozymes [14, 15], microsatellites [16, 17], PCR-RFLP [18, 19] and sequence polymorphism of genes and DNA fragments [20–22]. These studies have revealed the existence of Western and Eastern phylogroups [6, 19]. Individuals from both phylogroups have undergone natural and human-aided hybridization and this has produced hybrids that appear in natural water bodies as well as in cultured stocks along Europe.

The rapid development and application of sequencing technologies is now permitting researchers to discover thousands of SNPs at relatively low cost compared to the traditional Sanger sequencing method [23]. Transcriptome sequencing is considered a cost-effective strategy for discovering SNPs in non-model species. In fact, as a transcriptome is directly associated with functional regions in a genome, transcriptome-derived SNPs can be informative for adaptive variation [24–26] and they can be used not only for assessing population genetic structure, but also for genomic selection for traits of interest to aquaculture such as growth, sex determination or disease resistance (e.g. [27–29]). Given these advantages, SNPs derived from transcriptomes have been widely discovered and studied in many fish species [29–42].

The aim of this study was: to discover and validate transcriptome-derived SNPs in *T. tinca*, based on the strategy designed by Montes *et al*. and successfully applied in other fish species [38, 43]. The SNPs array was then used to disentangle the population genetic structure of two cultured tench breeds (Tabor and Hungarian), previously identified as stocks representing mixture of haplotypes out of both phylogroups [22].

Materials and methods

Ethics statement

The handling and usage of experimental fish in this study was done in accordance with the Czech Act. No 256/1992 Coll. as amended under supervision of the Institutional Animal Care and Use Committee (IACUC) of the University of South Bohemia (USB), Faculty of Fisheries and Protection of Waters (FFPW) in Vodňany. The USB FFPW has approval of the Ministry of Agriculture of the Czech Republic for handling and usage of experimental animal's ref. no. 16OZ15759/2013-17214. The presented study was included in the planned activities dealing with study of biodiversity, genetic, physiological and reproductive variability and performance of selected freshwater fish species. The experimental stock was reared under the common semi-intensive pond management conditions. The fish sacrificed for the study were euthanized in accordance with the Ordinance no. 419/2012 Coll. as amended. The fish were euthanized by blow into the head using a blunt object and bleeding. One of the co-authors was present during handling and processing the fish owned the certificate (no. 0135/2000-V3) which allows him to conduct and manage experiments involving animals according to the above mentioned act.

Sample collection

In the methodology followed for SNP discovery, two samplings (corresponding to the two sequencing approaches) were performed; one for transcriptome sequencing, and another for genome sequencing.

For transcriptome sequencing, 4 tench individuals (2 males and 2 females) were sampled. The sampled individuals belonged to two metabolic activities (summer season with 20˚C water temperature, and winter season with 4˚C water temperature) and two breeds (Hungarian and Tabor) cultured in Vodňany, Czech Republic since 1990's [44] (present Faculty of Fisheries and Production of Waters, University of South Bohemia in České Budějovice). The Tabor breed was established by collecting fish from ponds of a Czech county, and the Hungarian breed by introducing the tench from Hungary. To increase the homozygosity, inbreeding and gynogenesis within each breed were applied. Both breeds, containing approximately 120 adult individuals, have been maintained to date by intra-linear mating only for 6 generations. Previous studies on these fish have shown that both breeds have gene pools mixed of both Western and Eastern phylogroups [22, 45]. The transcriptome changes according to genes expressed. Expression of various genes depends on many inner and outer factors (e.g. fish age, health status, phase of reproductive cycle, weather, season—growing or wintering etc.). That is why we sampled fish in winter (no-growing season) and summer (growing season) in order to cover different genes expressed in mature 4-year old fish. Each fish was humanely sacrificed and two different tissues were collected- whole brain (without pituitary) and back muscle (approx. 1 g) and immediately frozen in liquid nitrogen, and stored at -80˚C until RNA extraction was performed. We had eight initial tissue samples, though two (brain in both cases) were not suitable for sequencing due to RIN values below 8. The remaining six samples (two of them in duplicate) were used for library construction and Illumina sequencing (S1 Table).

For genome sequencing, a total of ten tench individuals from six different locations were collected (S2 Table) in order to cover maximal available genetic diversity, including phylogroup origin of tench species. Samples were taken from the tench tissue collection of Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany and they represented populations throughout Neighbor-joining trees inferred from studies focused on genetic diversity of the growth hormone (GH) gene [22], microsatellites [17] and mitochondrial DNA [18].

RNA and DNA extraction

Total RNA was isolated using Qiazol lysis reagent (Qiagen). The isolated RNA was quantified with a Nanodrop 2000 (Thermo Scientific) and integrity of RNA (RIN) was determined using an Agilent 2100 Bioanalyzer (Agilent Technologies). Samples with RIN values above 8 were used for RNA sequencing, and used for library construction and Illumina sequencing. According to the RNA quality standards, six samples were sequenced (S1 Table).

Genomic DNA was isolated from muscle, fin or blood samples using the peqGOLD Tissue DNA Mini Kit (Peqlab Biotechnologie) following manufacturer instructions. The quantity and quality of DNA was measured with Qubit 2.0 Fluorometer and 0.8% agarose gel electrophoresis. The DNA samples with concentrations $\geq$ 50 ng/μl, 260/280 ratios of 1.8–2.0 and clear high molecular weight bands on the gel were used for genome sequencing. An equimolar amount of total DNA was then pooled for the library preparation.

Library construction and Illumina sequencing

A multiplex sequencing library was prepared by labeling each sample (six RNA samples, two of them replicated; and two DNA pools) with specific 10-mer barcoding oligonucleotides.

Transcriptomic and genomic libraries were sequenced in a single lane of Illumina HiSeq2000 and HiSeq2500 platforms, respectively. Sequencing reactions were performed separately for transcriptome and genome with paired-end 101 bp and 126 bp reads, respectively. Sequencing was performed at CNAG- Centre Nacional d'Anàlisi Genòmica, Barcelona, Spain. All sequence data have been submitted to NCBI's submission portal under the BioProject accession number PRJNA414567.

Genome size estimation

We estimated the genome size of *Tinca tinca* by means of the frequencies of the kmers in the DNA reads. Reads were processed with Jellyfish 2.2.10 [46] using the *count* subcommand with a kmer size of 25. The frequencies were computed using the *histo* subcommand. Finally, the genomic haploid length, along with the repetitive and unique contents and rate of heterozygosity, was computed using the GenomeScope web service [47].

Transcriptome *de novo* assembly and annotation

Raw RNA reads were processed in a strict four-step procedure in order to obtain a high-quality reference. First, adaptors and low-quality reads were removed with Trimmomatic v0.33 [48] by deleting the first 13 nucleotides of the read. Removal of adapters was done with the ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 parameters by setting a minimum mean PHRED quality value of 10, trailing bases with quality value at least 20, and a minimum read length of 31 bases. Second, contaminants indicated by the UniVec database were removed with SeqClean (https://sourceforge.net/projects/seqclean/). Third, Trimmomatic was run on Single End mode to remove low quality and excessively short reads with the following parameters: minlen:31 avgqual:10 minlen:31 trailing:19 minlen:31 tophred33. Finally, the paired-end structure of the reads was recovered with a custom script written in Python with help of the Biopython package [49].

After the transcriptome reads were trimmed, paired and unpaired high-quality reads (all RNASeq data) were assembled into contigs using Trinity v2.0.6 [50]. The resulting transcriptome was uploaded to NCBI Transcriptome Shotgun Assembly Sequence Database and it is available at GenBank with accession number GFZX00000000.1. Full implementation of assembly procedure is available at https://github.com/GenomicResources/ttin_assembly.

To measure the quality of the assembled transcriptome, we used a two-fold approach. First we backmapped (with Bowtie2) the trimmed reads against the generated reference to measure the fidelity of the assembly with respect to the reads. According to the authors of Trinity, transcriptomes with mapping rates above 80% are considered good assemblies. Second, we used BUSCO v3.0.2 [51, 52] to assess the quality of the assembly by searching for *Actynopterygii* Single Copy Orthologs (SCOs). The program searches the homology between our transcriptome and a set of precomputed proteins that are known to be conserved across the evolution of a large set of species, classifying them as SCOs, conserved but duplicated, fragmented, or missing.

Finally, TransDecoder v2.0.1 (https://transdecoder.github.io/) and Trinotate (http://trinotate.github.io/) were used for transcriptome annotation and generation of a tench proteome. Transdecoder is a pipeline that extracts the possible open reading frames (ORFs) from a raw transcriptome to predict if it has homology with BLAST [53] against a protein reference database like Swiss-Prot [54] (downloaded on August 2015), UniRef90 [55] (accessed on August 2015), or homology via Hidden Markov Models with HMMER [56] (retrieved on August 2015) by querying the Protein Families database (Pfam, [57]. Once ORFs are called and possible homologies to elements in the different databases are hypothesized, a proteome is generated.

The next step in the procedure is the annotation of both the transcriptome and the predicted proteome developed as described above with Trinotate. It consists of homology searches, as done in the TransDecoder step, with help of BLASTX, BLASTP and HMMER, to then make use of a database (downloaded on August 2015) containing annotations from Gene Ontology (The Gene Ontology Consortium, 2000), KEGG [58], and eggnog [59].

Chimeras and duplicated regions were filtered out from the assembled transcriptome with stringent filters. First, contigs were quantified with Kallisto [60] and those with zero counts were removed with help of the Sleuth R package [61]. Additionally, according to the generated proteome, contigs with no coding potential were removed. Finally, genes that produce two or more isoforms were deleted. These procedures were performed using custom scripts in Python, R (R Core Team 2016), SAMtools [62] and Snakemake [63]. Implementation of the annotation procedure is available online at https://github.com/GenomicResources/ttin_trinotate. The resulting filtered transcriptome was used in the following steps of intron-exon boundary (IEB) prediction and SNP discovery.

SNP calling and IEB prediction

Tench SNP calling was performed as described by [38]. Two parallel SNP calling approaches were performed by aligning transcriptome (T2T) and genome (G2T) trimmed reads to the filtered transcriptome. This alignment was performed with Bowtie2 in local mode [64]. In this pipeline, PCR duplicates from both transcriptome and genome reads were removed using the SAMtools *rmdup* command [62]. Subsequently, variants were called with SAMtools *mpileup* command [62]. In order to avoid false SNPs, a maximum contig depth of 200x was set to avoid both repetitive sequences and false positive local alignments; the minimum contig depth allowed for T2T variants was 8x and 20x in the case of G2T variants in order to remove transcripts with low coverage that could bias the SNP calling procedure; the minimum variant count allowed for T2T variants was 2 high quality (HQ) bases (i.e., the alternative base appears at least twice), and 3 HQ bases for G2T variants. This last filtering step requires the SNPs to have higher MAFs when coverage is lower. After applying all of these filters, only common variants present in both T2T and G2T SNP discovery approaches were considered as putative SNPs. The implementation of the transcriptome filtering and SNP calling procedures is available online at https://github.com/GenomicResources/ttin_snps.

It is well known that genotyping procedures (for PCR based technology like fluidigm) will fail if primers are spanning or otherwise close to intron-exon boundaries [65]. Therefore, the filtered transcriptome reference was *in silico* assessed to detect IEBs as described by [66]. This is done by mapping genomic reads to the transcriptome, and computing p-values for *change points*. These are locations in the transcriptome where one or more genomic reads do not map throughout their whole length but rather the mapping is initiated or terminated internally to the read. Locations with low p-values represent a surprising number of change points at that location, hence a likely IEB. Predicted IEBs are annotated and avoided during genotyping primer design.

SNP genotyping and validation

A total of 140 tench samples belonging to two breeds (Tabor, N = 66 and Hungarian, N = 74) were genotyped for selected subset of 96 candidate SNPs. Only one SNP per contig was chosen and selection was not biased to any gene family. As growth-related traits are of main importance in most cultured fish species and growth hormone (GH) gene might be associated with growth [6], the SNP array was within each breed also associated with GH gene genotype distinguishing alleles of Eastern or Western phylogroup haplotype. Assignment of an individual to

Eastern (E) only, Western (W) only or hybrid (H) GH gene genotype was performed using the sequence analysis of GH gene [22]. In the pure Western GH gene genotype the first GH gene fragment including polymorphic side 1 (PS1) and the second GH gene fragment including PS7 were 344 bp and 451 bp long, respectively, while the individuals of pure Eastern GH gene genotype had fragments of 341 bp and 455 bp in length, respectively. In hybrids, haplotypes of both phylogroups were observed (i.e. 341 and 344 for PS1 and 451 and 455 for PS7). Flanking sequences of a subset of SNPs selected for validation were used for primers and probe design according to Fluidigm Genotyping System requirements. After genotyping, SNPs were categorized as *no signal* (unamplified SNPs), *disperse* (call rate $< 80\%$), *monomorphic* (minor allele frequency, MAF < 0.01) and *psv* (paralogous sequence variant; all individuals are heterozygotes). For the conversion rate (proportion of all genotyped SNPs showing polymorphism), *no signal* and *disperse* SNPs were discarded, while only polymorphic SNPs (no *monomorphic*, neither *psv*) were used for the estimating the validation rate (proportion of polymorphic SNPs reliably scored in a sample of individuals). Polymorphic SNPs were uploaded to EBI's European Variation Archive under the study accession number PRJEB23783.

Population genetic structure

For each polymorphic SNP, minor allele frequency, and expected and observed heterozygosities (H_e and H_o, respectively) were estimated using the software package GeneClass2 [67]. Deviations from Hardy-Weinberg equilibrium (HWE) were evaluated for each *locus* using Fisher's exact test implemented in Genepop 4.0 [68] with 10,000 dememorizations, 100 batches and 5,000 iterations per batch.

To determine the genetic structure of tench individuals, genotype data were analyzed with Structure 2.3.4 software [69]. The number of clusters k was determined by comparing log-likelihood ratios in 10 runs for values of k between 1 and 10. Each run started with a burn-in period of 10,000 steps followed by 100,000 MCMC replicates. The optimal k was estimated as proposed by [69] and [70] and bar plots were generated using Pophelper v1.0.7 [71].

Based on this initial structure, the Bayesian likelihood method implemented in Bayescan 2.1 [72] was used to detect loci under natural selection (outlier loci). Bayescan was run with twenty pilot runs of 5,000 iterations, an additional burn-in of 50,000 iterations and prior odds of 10 for neutral model. Critical values were adjusted with a false discovery rate (FDR) procedure ($q<0.1$) [73]. Results of the outlier test were used to partition the SNP dataset into neutral and outlier loci; i.e., markers presumably under natural selection. Those loci resulting as outlier were removed from prospective analysis, regarding neutral variation, and annotations of the genomic regions including those loci were re-inspected.

Finally, neutral genetic differentiation and inbreeding were assessed. Neutral genetic differentiation was estimated with unbiased F_{ST} (distance matrix: pairwise difference) [74] using Arlequin v3.5 [75]. Inbreeding was estimated with F_{IS} [74] statistic using Fstat software [76]. The statistical significance of F_{ST} and F_{IS} was tested by 1,000 permutations for each pairwise comparison. In all cases with multiple comparisons, error rates were corrected using the sequential Bonferroni procedure [77].

Results

Transcriptome and genome sequencing

In total 32 million paired-end transcriptomic reads, with an average length of 101 bp, were sequenced (S3 Table). In the case of genome, 316 million genomic reads with a read length of 126 bp were generated, encompassing 154 million reads generated for Western pool (19.6 Gbp), and 162 million reads for Eastern pool (20.4 Gbp). GenomeScope estimated that the

Tinca tinca has a maximum genome size of 778,555,248 base pairs, where 599,234,146 base pairs (76.97%) constitute unique regions (S4 Table; S1 Fig). Overall, genome sequences constituted an estimated 51.58x coverage of the tench genome.

Transcriptome *de novo* assembly and annotation

Trimming of raw transcriptome reads did not result in a significant removal of reads, but 16% of nucleotides were discarded (S5 Table). The transcriptome *de novo* assembly consisted of 267,058 contigs (294.7 Mbp), which are the result of potentially 174,378 genes. The length of the assembled contigs ranged from 224 bp to 23,703 bp with an average length of 1,103 bp (S2 Fig).

Given the high number of sequences that Trinity yielded, we assessed the quality of our transcriptome by read mapping and by the contents of Single Copy Orthologs. On the one hand, the backmapping method achieved mapping success rates between 96.54% and 99.38% (S6 Table), suggesting therefore a good reconstruction of the *Tinca tinca* transcriptome. On the other hand, BUSCO reported that the transcriptome contains 85.9% of the Actinopterygii BUSCOs (where 40.4% are single copies), 6.7% are fragmented, and only 7.4% are completely missing (S7 Table). We conclude that given that even though we only sampled two tissues (muscle and brain) of *Tinca tinca*, this assembly is a good representation of the transcriptome.

According to the gene-isoform distribution in S3 Fig the distribution is skewed towards genes composed by one transcript. There are 10,705 genes of that composition (out of 174,378 genes, 86.42%, and out of 267,058 isoforms, 56.43%). The mean of the distribution is 1.53 transcripts per gene. As an extreme value, there is a gene (possibly a gene family) composed of 55 transcripts.

Regarding annotation, 89,832 transcripts were annotated (33.63%) as 126,187 proteins and 32,619 genes. From these, 64,676 transcripts (105,812 proteins and 9,295 genes) had a positive match to the UniRef90 database with *blastp* (S8 Table); similarly, 101,606 contigs (39,169 genes) were positively mapped with *blastx* (S9 Table). In both cases, top reference transcripts belonged to the same species: *Danio rerio*, *Astyanax mexicanus*, *Oncorhynchus mykiss*, *Oreochromis niloticus*, and *Ictalurus punctatus* (S4 and S5 Figs; S10 Table).

Overall, 67,953 contigs (77,626 proteins and 22,996 genes) were positively matched to 5,054 different protein domains according to the Pfam database (S6 Fig). The five most popular domains were: C2H2-type zinc finger (6.19%), Immunoglobulin domain (4.02%), Ankyrin repeat (3.22%), Leucine rich repeat (3.06%), and Zinc finger, C2H2 type (2.58%; S11 and S12 Tables).

According to the EggNOG database, 43,291 contigs (43,366 proteins and 14,714 genes) had a match against 3,338 different elements of the EggNOG database, including Serine Threonine protein kinase (7.63%), repeat-containing protein (3.03%), Zinc finger protein (2.95%), Ankyrin repeat (2.47%) and GTP-binding protein (1.27%) (S7 Fig and S13 Table).

Finally, Gene Ontology (GO) analysis showed 88,031 contigs (89,014 proteins and 30,345 genes). The highest number of GO terms was assigned to biological processes (48.63%) followed by molecular functions (29.66%) while cellular component has the least assigned terms (21.70%; S8 Fig). The three most commonly assigned GO terms in biological process category were genes involved in *Transcription*, *DNA-templated* (2.03%), *Regulation of Transcription*, *DNA-templated* (1.38%) and *Signal Transduction* (0.73%). In the molecular function ontology, *ATP binding* (5.77%), *Metal ion binding* (5.32%), *Zinc ion binding* (4.08%) and *DNA binding* (4.06%) were the most represented terms. The three major assigned GO terms for cellular component were nucleus (10.51%), cytoplasm (10.35%) and integral components of the membrane (7.26%; S9–S12 Figs; S14 and S15 Tables).

Table 1. Descriptive statistics of G2T, T2T and common discovered SNPs.

	G2T	T2T	Common
Contigs with SNPs	15,593	13,721	16,263
Number of contigs in filtered assembly	18,479	18,479	18,479
Transcripts with SNPs (%)	84.38	74.25	88.01
SNPs number	131,188	98,869	169,643
Assembly size (bp)	20,316,163	20,316,163	20,316,164
Mean mutation rate (SNPs/bp)	0.006	0.005	0.008
SNPs per transcript	8.41	0.14	0.10

https://doi.org/10.1371/journal.pone.0213992.t001

SNP discovery and validation

According to kallisto, a total of 262,801 contigs (out of 267,058) had an expression value above zero transcripts per million (TPM). Therefore 98.41% of the original assembly remained valid for further analysis. From those, 89,832 contigs were identified as having no coding potential and were discarded. Finally, contigs representing more than one isoform were also removed. After all these filters, the transcriptome was reduced to 18,479 contigs spanning 20.32 Mbp.

The filtered transcriptome was used as reference for mapping genome (G2T) and transcriptome (T2T) trimmed reads. The trimming process did not significantly decrease the number of transcriptome or genome reads (S16 and S17 Tables). The mapping process resulted in 19.51% of genomic reads and 22.63% of transcriptome reads assigned to the filtered transcriptome (S18 Table). From these mappings, a total of 131,188 G2T SNPs were called in 15,593 transcripts (8.41 G2T SNPs/transcript; Table 1), and 98,869 T2T SNPs were called in 13,721 transcripts (7.21 T2T SNPs/transcript). Together, G2T and T2T called 169,643 SNPs in 16,263 transcripts, but only 60,414 SNPs in 11,769 transcripts (5.13 SNPs/transcript) were common to both sets. These 60,414 SNPs represented the final set of putative SNPs discovered in the tench transcriptome.

Regarding IEB avoidance, 4,091 transcripts out of 18,479 were signaled as not having multi-mapped reads (those that map to more than one transcript); and a total of 2,937 transcripts contained one or more predicted IEB. A total of 16,764 IEBs were predicted (on average 5.70 predicted IEB per transcript). These predicted IEBs were annotated and avoided during genotyping primer design.

A set of 96 SNPs was selected based on IEB prediction analysis for validation and genotyping on Fluidigm Genotyping System. From the 96 SNPs that were genotyped, 4 (4.17%) were categorized as *no signal*, while the remaining 92 SNPs were *polymorphic* with >80% call rate. Therefore, conversion and validation rates of 95.83% were achieved.

Population genetics

Mean H_o and H_e for the Hungarian breed were 0.508 and 0.460, respectively. Similar levels of H_o (0.455) and H_e (0.458) were found in the Tabor breed. Tests of deviation from HWE for each locus revealed no significant departure from HWE after sequential Bonferroni correction. The STRUCTURE analysis evidenced population structure with K = 2 (Evanno method; Fig 1A), and K = 3 (Pritchard method; Fig 1B) being the most likely number of clusters. The average of the mean posterior probability (LnP(D)) estimated from 10 independent runs on K = 2 and K = 3 was -16533.7 and -16176.1, respectively. These clusters clearly indicate the differences between the two breeds, but not between the GH gene genotypes (Fig 1).

A total of six SNPs were detected as being under diversifying selection (positive alpha values); this is, they show extremely different allele frequencies in the two breeds. These outlier

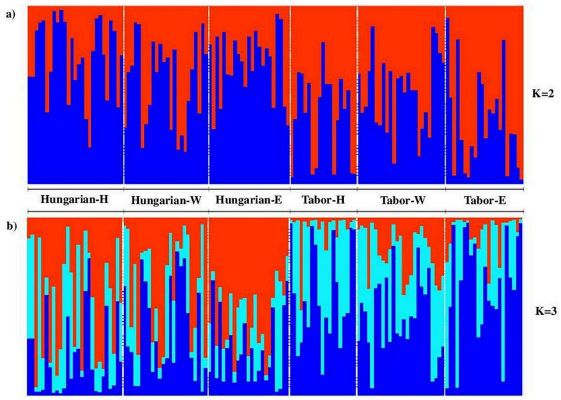


Fig 1. Results from STRUCTURE analysis for K = 2 (a) and K = 3 (b). Individuals corresponding to each breed (Hungarian, Tabor) and GH gene phylogroup genotype (H: Hybrid; W: Western; E: Eastern) are separated with vertical white bars.

https://doi.org/10.1371/journal.pone.0213992.g001

SNPs were located in the following genes: MRPL32 (39S ribosomal protein L32, mitochondrial), CENPF (Centromere Protein F), GRM1 (Glutamate Metabotropic Receptor 1), SPRY4 (Protein sprouty homolog 4), TRIP4 (Thyroid Hormone Receptor Interactor 4) and CN080 (Uncharacterized protein C14orf80 homolog) (Table 2). Of these six SNPs, all were found to be synonymous mutations, except interestingly for the SNP within *Activating signal cointegrator 1* (see Table 2) which encodes for either a Val (hydrophobic amino acid) or a Ser (polar

Table 2. Annotation of selected loci based top BLAST hit and GO ontology.

Locus ID	Genomic BLAST Hit	GO ID	e-value	Gene function
TR107177|c0_g1_i1	Sprouty homolog 4-like	GO:001602 GO:0021594GO:0030097GO:0040037GO:004874 GO:0070373	0.0E0	P: Negative regulation of fibroblast growth factor receptor signaling pathway; P: Rhombomere formation; P: Skeletal muscle fiber development; P: Hemopoiesis; P: Negative regulation of ERK1 and ERK2 cascade; C: Membrane
TR57930|c0_g1_i1	Centromere F	GO:0008134GO:0042803GO:0045502	0.0E0	F: Protein homodimerization activity; F: Transcription factor binding; F: Dynein binding
TR48380|c0_g1_i1	39S ribosomal L32, mitochondrial	GO:0005743GO:0005762GO:0003735GO:0016787GO:0006412	2.2 E-105	F: Structural constituent of ribosome; C: Mitochondrial large ribosomal subunit; C: Mitochondrial inner membrane; F: Hydrolase activity; P: Translation
TR71953|c0_g1_i1	Metabotropic glutamate receptor 1-like isoform X1	GO:0016020GO:0004871GO:0007165	1.1E-153	P: Signal transduction; C: Membrane; F: Signal transducer activity
TR96558|c0_g1_i2	Activating signal cointegrator 1	GO:0005634GO:0003713GO:0008270GO:0006366GO:0045893	0.0E0	C: Nucleus; F: Zinc ion binding; F: Transcription coactivator activity; P: Positive regulation of transcription, DNA-templated; P: Transcription from RNA polymerase II promoter
TR56671|c0_g1_i1	Uncharacterized protein C14orf80 homolog isoform X1	-	0.0E0	-

https://doi.org/10.1371/journal.pone.0213992.t002

amino acid). Since this substitutes a polar amino acid for a hydrophobic one, this SNP may lead to a change in protein function and should be further explored. Functional annotation revealed that most of these genes encoded proteins involved in transcription and translational regulation and structural organization of ribosome and mitochondria. Apart from these, the annotated gene Sprouty homolog 4-like was found to be involved in regulation of fibroblast growth and skeletal muscle fiber development, suggesting that the studied tench breeds might be adapted to different environments that affect growth related genes.

After removing the 6 outlier SNPs, a set composed of 86 SNP markers was used for studying neutral genetic differentiation and inbreeding. Pairwise F_{ST} estimates, within each breed, among E and W phylogroups and EW hybrid (H) were not significant; in contrast, all F_{ST} values were significant when pairwise comparisons between the two breeds were tested (Table 3). Overall, F_{ST} value between the two breeds was low but significant (F_{ST} = 0.0450, p-value < 0.0001). Additionally, F_{IS} within each breed was not significant, indicating homogeneity within breeds. In summary, individuals within breeds show homogeneous allele frequencies without regard to GH gene genotype, whereas individuals of the two breeds (even if they both are a mixture of E and W phylogroup haplotypes) are genetically different. Genotyping results of all 92 SNPs markers have also been included in the S19 Table.

Discussion

The major challenge of transcriptome-derived SNPs is marker "drop-out" during the validation step; the most significant factor is if a SNP spans an IEB. For instance, 64% of genotyping failures have been reported in EST-derived SNPs in catfish due to the proximity of SNPs to IEB [65]. The most evident cause for such genotyping failure is the presence of priming site at SNPs loci leading to non-base pairing of primers or expected amplification product is too large for amplification due to presence of intron between priming sites. Therefore, the key for successful SNP validation is avoidance of IEBs. In this study, the approach devised by [66] and applied successfully to European anchovy [38] was used to avoid the problem related to IEBs. In this method, the assembled transcript sequences were aligned to genome sequences of tench to identify the IEB. By selecting the SNPs not spanning an IEB, we obtained the highest conversion and validation rates of transcriptome-derived SNPs obtained to date for a non-model species.

In this study, using the validated SNPs we have demonstrated that the two tench breeds show low but significant genetic differentiation, even with their similar genetic structure concerning their phylogroup based gene pool. The ancestral populations that formed the two tench phylogroups separated about 0.064 to 1.6 million years ago as revealed from 1.6% sequence divergence of cytochrome b mitochondrial gene [21]. The western (W) and Eastern

Table 3. Pairwise F_{ST} (below diagonal) and p-values (above diagonal) among tench breeds (Hungarian, Tabor) and GH gene phylogroups genotype (H: Hybrid; W: Western; E: Eastern).

	Hungarian -H	Hungarian -W	Hungarian -E	Tabor-H	Tabor-W	Tabor-E
Hungarian-H	-	0.2022	0.1592	0.0000	0.0000	0.0000
Hungarian -W	0.0012	-	0.0429	0.0379	0.0000	0.0000
Hungarian -E	0.0025	0.0083	-	0.0787	0.0504	0.0000
Tabor-H	0.0619*	0.0000	0.0000	-	0.1973	0.3936
Tabor-W	0.0399*	0.0274*	0.0000	0.00538	-	0.0049
Tabor-E	0.0579*	0.0318*	0.0687*	0.00150	0.0218	-

* significant value

https://doi.org/10.1371/journal.pone.0213992.t003

(E) phylogroup significantly differs also in sequences of nuclear DNA e.g. the second intron of the actin gene, an intron of the gene coding for the ATP synthase β subunit, the first intron of the gene coding for the S7 ribosomal protein [21] and GH gene [6]. Due to the long history of tench phylogroup separation and individual evolution it is expected that the phylogroups would differ significantly also in physiological and biological functions resulting from nucleotide polymorphisms of functional genes. Therefore, our transcriptome-derived SNP array could be used for screening tench populations that still contain haplotypes of pure Western and pure Eastern phylogroup or F1 hybrid generation between pure W and E tench populations. Unfortunately, tench populations that bear pure Western haplotypes are very scarce or even absent [21] and we did not have such population in our collection. The Hungarian and Tabor breeds are, after several generations of mating fish with haplotypes of both phylogroups, a mosaic of both phylogroups due to free combination of chromosomes, crossing overs between homologous chromosomes and other possible processes that appear during formation of gametes. Based on F_{ST} values inferred from 86 SNPs it can be indirectly assumed that the SNPs genotypes were not significantly different for fish having Eastern, Western or hybrid GH gene genotype [22] within both Tabor and Hungarian breed. If the rate of phylogroup introgression within breeds were low, the degree of differentiation among fish displaying different GH gene genotype would be expected due to previously mentioned divergence between phylogroups in other genetic markers. On the other hand, significantly different F_{ST} values were observed between the two breeds with no matter to what GH gene genotype the fish belonged. The within-breed gene flow is corroborated by previous studies that show no negative fitness consequences derived from two phylogroup-mixed tench populations under cultured conditions [78]. In summary, six generations of within-breed isolated reproduction under cultured conditions allowed breed identity determination using the transcriptome-derived SNP array.

Moreover, apart from neutral levels of genetic differentiation, the SNPs in this study are transcriptome-derived markers and their variation in genes is informative for differential selection or adaptation in each breed. In this study, high allelic differentiation between both breeds was observed in growth-related genes, which might point to differential natural and human-affected selection, breeding and evolutionary history of Hungarian and Tabor tench breeds and/or stocks they were established from. Taking into account that the sequence of the GH (growth hormone) gene has 0.8% divergence in both tench phylogroups [6], we propose the following hypothesis: adaptive differences between breeds arise from differential composition of individuals from each phylogroup in each breed, giving to Hungarian and Tabor breeds different weight to their adaptation affecting growth related genes. However, further studies with protein sequencing of genes under selection are needed to corroborate the hypothesis presented here, as most of the SNPs found in the genes under selection have arisen due to synonymous mutations and will not lead to a change in the protein configuration. Insignificant association between GH gene genotype and SNP array also indicates that there is no linkage between our SNPs and the GH gene. However, this result does not say anything about association of these two markers to growth-related traits. It seems that effects of SNP array and GH gene genotype polymorphism on the growth-related traits will be (if any) independent of each other.

This study represents the first large-scale sequencing effort for SNP discovery and validation in tench. Although restriction-site associated DNA sequencing (RADseq) or double digest RADseq (ddRADseq) can generate large data set, SNPs derived from these approaches mostly fall into non-coding or unknown regions. Transcriptome derived SNPs are directly associated with functional regions in the genome and can give more information for 92 SNPs in coding region than hundreds or thousands of SNPs derived from non-coding or (not identified) regions. The validated SNPs can be used in further genetic studies for finding genes and/or DNA sequences associated with trait of importance.

Conclusions

The SNP discovery approach followed in the present study was developed for transcriptome-derived SNP discovery in European anchovy [38], and Atlantic mackerel [43] with successful conversion and validation rates. This approach can be used to discover large number of transcriptome-derived SNPs in any non-model species. In addition, our approach identifies SNPs in the transcriptome: these SNPs can be annotated and in some cases, as evidenced here, they are under natural selection. We showed that the SNPs array in tench is strong enough to distinguish tench breeds and that it might be useful for studies focused on searching the range of associations between DNA sequence and traits of importance. Overall, it was verified that transcriptome-derived SNPs may informs us not only about neutral genetic differentiation and population genetic structure (e.g. [37, 39]), but also about the functional role of the differences observed between populations or ecotypes

Supporting information

S1 Fig. GenomeScope profile.
(TIFF)

S2 Fig. Transcript-length distribution.
(TIF)

S3 Fig. Gene—Transcript distribution after TransDecoder prediction.
(TIF)

S4 Fig. Blastp hits distribution by organism. Organism is encoded by 5 letters (ASTMX *Astyanax mexicanus*, DANRE *Danio rerio*; ICTPU *Ictalurus punctatus*; LEPOC; ONCMY *Oncorryncus mykiss*; ORENI *Oreochromis niloticus*; POEFO; TAKRU *Takifugu rubripes*; TETNG; XIPMA).
(TIF)

S5 Fig. Blastx hits distribution by organism. Organism is encoded by 5 letters (ASTMX *Astyanax mexicanus*, DANRE *Danio rerio*; ICTPU *Ictalurus punctatus*; LEPOC; ONCMY *Oncorryncus mykiss*; ORENI *Oreochromis niloticus*; POEFO; TAKRU *Takifugu rubripes*; TETNG; XIPMA).
(TIF)

S6 Fig. Frequency distribution of Pfam domains (top 20).
(TIF)

S7 Fig. EggNOG distribution by ID (top 20).
(TIF)

S8 Fig. Gene Ontology summary.
(TIF)

S9 Fig. Joint GO level 2 distribution (top 30).
(TIF)

S10 Fig. Distribution of the Biological Process Gene Ontology terms.
(TIF)

S11 Fig. Distribution of the Cellular Component Gene Ontology terms.
(TIF)

S12 Fig. Distribution of the Molecular Function Gene Ontology terms.
(TIF)

S1 Table. Details of samples included in transcriptome sequencing.
(XLSX)

S2 Table. Details of samples included in genome sequencing.
(XLSX)

S3 Table. Transcriptome and genome sequencing results.
(XLSX)

S4 Table. Genome size estimation.
(XLSX)

S5 Table. Results obtained from the trimming performed in order to do transcriptome assembly.
(XLSX)

S6 Table. Backmapping of the reads.
(XLSX)

S7 Table. BUSCO.
(XLSX)

S8 Table. UniRef90 results from proteome querying (blastp).
(XLSX)

S9 Table. UniRef90 results from transcriptome querying (blastx).
(XLSX)

S10 Table. Summary of blastp results against UniRef90 by organism. Organism is encoded by 5 letters (ASTMX *Astyanax mexicanus*; DANRE *Danio rerio*; ICTPU *Ictalurus punctatus*; LEPOC; ONCMY *Oncorryncus mykiss*; ORENI *Oreochromis niloticus*; POEFO; TAKRU *Takifugu rubripes*; TETNG; XIPMA).
(XLSX)

S11 Table. Pfam hits.
(XLSX)

S12 Table. Pfam hits summarized by domain.
(XLSX)

S13 Table. EggNOG results.
(XLSX)

S14 Table. Summary of Gene Ontology results (level 1).
(XLSX)

S15 Table. Gene Ontology annotation of the transcriptome and proteome.
(XLSX)

S16 Table. Trimming for SNP discovery (DNA).
(XLSX)

S17 Table. Trimming for SNP discovery (RNA).
(XLSX)

S18 Table. Mapping.
(XLSX)

S19 Table. Genotyping results of all the 92 SNPs markers.
(XLSX)

Acknowledgments

The authors are thankful for the technical and human support provided by the Sequencing and Genotyping SGIker unit of UPV/EHU.

Author Contributions

Conceptualization: Martin Kocour, Andone Estonba.

Data curation: Girish Kumar.

Formal analysis: Girish Kumar, Jorge Langa, Iratxe Montes, Darrell Conklin.

Methodology: Girish Kumar.

Resources: Klaus Kohlmann.

Software: Girish Kumar, Jorge Langa, Darrell Conklin.

Supervision: Martin Kocour, Andone Estonba.

Writing – original draft: Girish Kumar.

Writing – review & editing: Girish Kumar, Jorge Langa, Iratxe Montes, Darrell Conklin, Martin Kocour, Andone Estonba.

References

1. Linhart O, Rodina M, Kocour M, Gela D. Insemination, fertilization and gamete management in tench, *Tinca tinca* (L.). Aquacult Int. 2006; 14(1–2):61–73.
2. Wolnicki J, Kaminski R, Sikorska J. Combined effects of water temperature and daily food availability period on the growth and survival of tench (*Tinca tinca*) larvae. Aquac Res. 2017; 48(7):3809–16.
3. Welcomme RL. International introductions of inland aquatic species. Rome: Food and Agriculture Organization of the United Nations; 1988. 318 p.
4. Kocour M, Gela D, Rodina M, Flajshans M. Performance of different tench, *Tinca tinca* (L.), groups under semi-intensive pond conditions: it is worth establishing a coordinated breeding program. Rev Fish Biol Fisher. 2010; 20(3):345–55.
5. Wang JX, Min WQ, Guan M, Gong LJ, Ren J, Huang Z, et al. Tench farming in China: present status and future prospects. Aquacult Int. 2006; 14(1–2):205–8.
6. Kocour M, Kohlmann K. Growth hormone gene polymorphisms in tench, *Tinca tinca* L. Aquaculture. 2011; 310(3–4):298–304.
7. FAO. Fishery Statistical Collections, Global aquaculture production 2017 April 20, 2017.
8. Flajshans M, Gela D, Kocour M, Buchtova H, Rodina M, Psenicka M, et al. A review on the potential of triploid tench for aquaculture. Rev Fish Biol Fisher. 2010; 20(3):317–29.
9. Kvasnicka P, Flajshans M, Rab P, Linhart O. Inheritance studies of blue and golden varieties of tench (Pisces: *Tinca tinca* L.). Journal of Heredity. 1998; 89(6):553–6.
10. Svobodova Z, Kolarova J. A review of the diseases and contaminant related mortalities of tench (*Tinca tinca* L.). Vet Med-Czech. 2004; 49(1):19–34.
11. Chen WJ, Mayden RL. Molecular systematics of the Cyprinoidea (Teleostei: Cypriniformes), the world's largest clade of freshwater fishes: further evidence from six nuclear genes. Molecular phylogenetics and evolution. 2009; 52(2):544–9. https://doi.org/10.1016/j.ympev.2009.01.006 PMID: 19489125
12. Arslan A, Taki FN. C-banded karyotype and nucleolar organizer regions of *Tinca tinca* (Cyprinidae) from Turkey. Caryologia. 2012; 65(3):246–9.

13. Leggatt RA, Iwama GK. Occurrence of polyploidy in the fishes. Rev Fish Biol Fisher. 2003; 13(3):237–46.
14. Šlechtová V, Šlechtá V., Valenta M. Genetic protein variability in tench (*Tinca tinca* L.) stocks in Czech Republic. Polish Archives of Hydrobiology. 1995; 42:133–40.
15. Kohlmann K, Kersten P. Enzyme variability in a wild population of tench (*Tinca tinca*). Polish Archives of Hydrobiology. 1998; 45:303–10.
16. Kohlmann K, Kersten P, Flajshans M. Comparison of microsatellite variability in wild and cultured tench (*Tinca tinca*). Aquaculture. 2007; 272:S147–S51.
17. Kohlmann K, Kersten P, Panicz R, Memis D, Flajshans M. Genetic variability and differentiation of wild and cultured tench populations inferred from microsatellite loci. Rev Fish Biol Fisher. 2010; 20(3):279–88.
18. Lo Presti R, Kohlmann K, Kersten P, Gasco L, Lisa C, Di Stasio L. Genetic variability in tench (*Tinca tinca* L.) as revealed by PCR-RFLP analysis of mitochondrial DNA. Ital J Anim Sci. 2012; 11(1).
19. Lajbner Z, Kotlik P. PCR-RFLP assays to distinguish the Western and Eastern phylogroups in wild and cultured tench *Tinca tinca*. Molecular ecology resources. 2011; 11(2):374–7. https://doi.org/10.1111/j.1755-0998.2010.02914.x PMID: 21429147
20. Lo Presti R, Kohlmann K, Kersten P, Lisa C, Di Stasio L. Sequence variability at the mitochondrial ND1, ND6, cyt b and D-loop segments in tench (*Tinca tinca* L.). J Appl Ichthyol. 2014; 30:15–21.
21. Lajbner Z, Linhart O, Kotlik P. Human-aided dispersal has altered but not erased the phylogeography of the tench. Evolutionary applications. 2011; 4(4):545–61. https://doi.org/10.1111/j.1752-4571.2010.00174.x PMID: 25568004
22. Kocour M, Kohlmann K. Distribution of five growth hormone gene haplogroups in wild and cultured tench, *Tinca tinca* L., populations. J Appl Ichthyol. 2014; 30:22–8.
23. Metzker ML. Sequencing technologies—the next generation. Nature reviews Genetics. 2010; 11(1):31–46. https://doi.org/10.1038/nrg2626 PMID: 19997069
24. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nature reviews Genetics. 2003; 4(12):981–94. https://doi.org/10.1038/nrg1226 PMID: 14631358
25. Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004; 13(4):969–80. PMID: 15012769
26. Morin PA, Luikart G, Wayne RK, Grp SW. SNPs in ecology, evolution and conservation. Trends Ecol Evol. 2004; 19(4):208–16.
27. Bester-Van Der Merwe A, Blaauw S, Du Plessis J, Roodt-Wilding R. Transcriptome-wide single nucleotide polymorphisms (SNPs) for abalone (Haliotis midae): validation and application using GoldenGate medium-throughput genotyping assays. International journal of molecular sciences. 2013; 14(9):19341–60. https://doi.org/10.3390/ijms140919341 PMID: 24065109
28. Li S, Liu H, Bai J, Zhu X. Transcriptome assembly and identification of genes and SNPs associated with growth traits in largemouth bass (Micropterus salmoides). Genetica. 2017; 145(2):175–87. https://doi.org/10.1007/s10709-017-9956-z PMID: 28204905
29. Liao Z, Wan Q, Shang X, Su J. Large-scale SNP screenings identify markers linked with GCRV resistant traits through transcriptomes of individuals and cell lines in *Ctenopharyngodon idella*. Sci Rep. 2017; 7(1):1184. https://doi.org/10.1038/s41598-017-01338-7 PMID: 28446772
30. Ogden R. Unlocking the potential of genomic technologies for wildlife forensics. Molecular ecology resources. 2011; 11 Suppl 1:109–16.
31. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, et al. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular ecology resources. 2011; 11 Suppl 1:123–36.
32. Helyar SJ, Limborg MT, Bekkevold D, Babbucci M, van Houdt J, Maes GE, et al. SNP discovery using Next Generation Transcriptomic Sequencing in Atlantic herring (Clupea harengus). Plos One. 2012; 7(8):e42089. https://doi.org/10.1371/journal.pone.0042089 PMID: 22879907
33. Lamichhaney S, Martinez Barrio A, Rafati N, Sundstrom G, Rubin CJ, Gilbert ER, et al. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proc Natl Acad Sci U S A. 2012; 109(47):19345–50. https://doi.org/10.1073/pnas.1216128109 PMID: 23134729
34. Xu J, Ji P, Zhao Z, Zhang Y, Feng J, Wang J, et al. Genome-wide SNP discovery from transcriptome of four common carp strains. Plos One. 2012; 7(10):e48140. https://doi.org/10.1371/journal.pone.0048140 PMID: 23110192
35. Zarraonaindia I, Iriondo M, Albaina A, Pardo MA, Manzano C, Grant WS, et al. Multiple SNP markers reveal fine-scale population and deep phylogeographic structure in European anchovy (*Engraulis*

encrasicolus L.). Plos One. 2012; 7(7):e42201. https://doi.org/10.1371/journal.pone.0042201 PMID: 22860082

36. Hess JE, Campbell NR, Close DA, Docker MF, Narum SR. Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. Mol Ecol. 2013; 22(11):2898–916. https://doi.org/10.1111/mec.12150 PMID: 23205767

37. Montes I, Iriondo M, Manzano C, Santos M, Conklin D, Carvalho GR, et al. No loss of genetic diversity in the exploited and recently collapsed population of Bay of Biscay anchovy (*Engraulis encrasicolus*, L.). Mar Biol. 2016; 163(5).

38. Montes I, Conklin D, Albaina A, Creer S, Carvalho GR, Santos M, et al. SNP discovery in European anchovy (*Engraulis encrasicolus*, L) by high-throughput transcriptome and genome sequencing. Plos One. 2013; 8(8):e70051. https://doi.org/10.1371/journal.pone.0070051 PMID: 23936375

39. Montes I, Zarraonaindia I, Iriondo M, Grant WS, Manzano C, Cotano U, et al. Transcriptome analysis deciphers evolutionary mechanisms underlying genetic differentiation between coastal and offshore anchovy populations in the Bay of Biscay. Mar Biol. 2016; 163(10).

40. Laconcha U, Iriondo M, Arrizabalaga H, Manzano C, Markaide P, Montes I, et al. New Nuclear SNP Markers Unravel the Genetic Structure and Effective Population Size of Albacore Tuna (T*hunnus alalunga*). Plos One. 2015; 10(6):e0128247. https://doi.org/10.1371/journal.pone.0128247 PMID: 26090851

41. Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang H, et al. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife. 2016;5.

42. Robledo D, Rubiolo JA, Cabaleiro S, Martinez P, Bouza C. Differential gene expression and SNP association between fast- and slow-growing turbot (*Scophthalmus maximus*). Sci Rep. 2017; 7(1):12105. https://doi.org/10.1038/s41598-017-12459-4 PMID: 28935875

43. Alvarez P, Arthofer W, Coelho MM, Conklin D, Estonba A, Grosso AR, et al. Genomic Resources Notes Accepted 1 June 2015–31 July 2015. Molecular ecology resources. 2015; 15(6):1510–2. https://doi.org/10.1111/1755-0998.12454 PMID: 26452560

44. Flajshans M, Linhart O, Slechtova V, Slechta V. Genetic resources of commercially important fish species in the Czech Republic: present state and future strategy. Aquaculture. 1999; 173(1–4):471–83.

45. Lajbner Z, Kohlmann K, Linhart O, Kotlik P. Lack of reproductive isolation between the Western and Eastern phylogroups of the tench. Rev Fish Biol Fisher. 2010; 20(3):289–300.

46. Marçais G, Kinsgford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics. 2011; 27(6):764–770. https://doi.org/10.1093/bioinformatics/btr011 PMID: 21217122

47. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017; 33(14):2202–2204. https://doi.org/10.1093/bioinformatics/btx153 PMID: 28369201

48. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England). 2014; 30(15):2114–20.

49. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England). 2009; 25(11):1422–3.

50. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011; 29(7):644–52. https://doi.org/10.1038/nbt.1883 PMID: 21572440

51. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 2018; 35(3):543–548.

52. Felipe A. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717

53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

54. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research. 2003; 31(1):365–70. PMID: 12520024

55. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics (Oxford, England). 2007; 23(10):1282–8.

56. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011; 39(Web Server issue):W29–37. https://doi.org/10.1093/nar/gkr367 PMID: 21593126

57. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2014; 42(Database issue):D222–30. https://doi.org/10.1093/nar/gkt1223 PMID: 24288371

58. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research. 2012; 40(Database issue):D109–14. https://doi.org/10.1093/nar/gkr988 PMID: 22080510

59. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic acids research. 2012; 40(Database issue):D284–9. https://doi.org/10.1093/nar/gkr1060 PMID: 22096231

60. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology. 2016; 34(5):525–7. https://doi.org/10.1038/nbt.3519 PMID: 27043002

61. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nature methods. 2017; 14(7):687–90. https://doi.org/10.1038/nmeth.4324 PMID: 28581496

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009; 25(16):2078–9.

63. Koster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics (Oxford, England). 2012; 28(19):2520–2.

64. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9 (4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

65. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, et al. Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics. 2008; 9:450. https://doi.org/10.1186/1471-2164-9-450 PMID: 18826589

66. Conklin D, Montes, I., Albaina, A., Estonba, A. Improved conversion rates for SNP genotyping of non-model organisms. International Work-Conference on Bioinformatics and Biomedical Engineering; Granada, Spain 2013. p. 127–34.

67. Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A. GENECLASS2: a software for genetic assignment and first-generation migrant detection. The Journal of heredity. 2004; 95(6):536–9. https://doi.org/10.1093/jhered/esh074 PMID: 15475402

68. Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular ecology resources. 2008; 8(1):103–6. https://doi.org/10.1111/j.1471-8286.2007.01931.x PMID: 21585727

69. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945–59. PMID: 10835412

70. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005; 14(8):2611–20. https://doi.org/10.1111/j.1365-294X.2005.02553.x PMID: 15969739

71. Francis RM. pophelper: an R package and web app to analyse and visualize population structure. Molecular ecology resources. 2017; 17(1):27–32. https://doi.org/10.1111/1755-0998.12509 PMID: 26850166

72. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008; 180(2):977–93. https://doi.org/10.1534/genetics.108.092221 PMID: 18780740

73. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met. 1995; 57(1):289–300.

74. Weir BS, Cockerham CC. Estimating f-statistics for the analysis of population structure. Evolution; international journal of organic evolution. 1984; 38(6):1358–70. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x PMID: 28563791

75. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources. 2010; 10(3):564–7. https://doi.org/10.1111/j.1755-0998.2010.02847.x PMID: 21565059

76. Petit E, Balloux F, Goudet J. Sex-biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters. Evolution; international journal of organic evolution. 2001; 55(3):635–40. PMID: 11327171

77. Rice WR. Analyzing tables of statistical tests. Evolution; international journal of organic evolution. 1989; 43(1):223–5. https://doi.org/10.1111/j.1558-5646.1989.tb04220.x PMID: 28568501

78. Kumar G, Kohlmann, K., Gela, D., Kocour, M. Phylogroup origin of Tench Tinca tinca L. has no effects on main performance parameters. Aquaculture Europe-14; San-Sebastian, Spain 2014.

Appendix B

Article 2. EXFI

Langa, J., Estonba, A., & Conklin, D. (2020). EXFI: Exon and splice graph prediction without a reference genome. *Ecology and Evolution*, *10*(16), 8880–8893. https://doi.org/https://doi.org/10.1002/ece3.6587.

TABLE B.1: Quality Metrics for Ecology and Evolution in 2020.

Ecology and Evolution 2020 - Web of Science	
Category	Ecology
Impact Factor	2.912
Rank	70/166
Quantile	Q2

Ecology and Evolution 2020 - Scopus	
Category	Environmental Science - Ecology
CiteScore	4.1
Rank	80/400
Quantile	Q1

Received: 11 March 2020 | Revised: 3 June 2020 | Accepted: 8 June 2020

DOI: 10.1002/ece3.6587

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

EXFI: Exon and splice graph prediction without a reference genome

Jorge Langa[1] | **Andone Estonba**[1] | **Darrell Conklin**[2,3]

[1]Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country, Leioa, Spain

[2]Department of Computer Science and Artificial Intelligence, Faculty of Computer Science, University of the Basque Country UPV/EHU, San Sebastián, Spain

[3]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Correspondence
Jorge Langa, Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country, Barrio Sarriena S/N, Leioa, Pais Vasco, Spain.
Email: jorgeeliseo.langa@ehu.eus

Funding information
Basque Government, Grant/Award Number: predoctoral grant PRE_2017_2_0169 and grant IT558-10

Abstract

For population genetic studies in nonmodel organisms, it is important to use every single source of genomic information. This paper presents EXFI, a Python pipeline that predicts the splice graph and exon sequences using an assembled transcriptome and raw whole-genome sequencing reads. The main algorithm uses Bloom filters to remove reads that are not part of the transcriptome, to predict the intron–exon boundaries, to then proceed to call exons from the assembly, and to generate the underlying splice graph. The results are returned in GFA1 format, which encodes both the predicted exon sequences and how they are connected to form transcripts. EXFI is written in Python, tested on Linux platforms, and the source code is available under the MIT License at https://github.com/jlanga/exfi.

KEYWORDS

exome sequencing, exon, sequence capture, SNP discovery, splice graph, transcriptome

1 | INTRODUCTION

In the last decade, high-throughput sequencing technologies have enabled biologists to unravel the genetic code on a massive scale and at an unprecedented rate. However, sequencing and assembling whole genomes of nonmodel species is still not practical. Therefore, alternative approaches are needed to capture genetic variation. One approach commonly used in the context of population genetics is restriction site-associated DNA sequencing (RAD-Seq; Baird et al., 2008), which returns polymorphic markers at random loci across the entire genome. Posterior enhancements, such as RAD-Seq followed by sequence capture (Rapture; Ali et al., 2016), have been recently proposed as an efficient and cost-effective approach for genotyping thousands of samples and loci simultaneously (Meek & Larson, 2019).

Another successfully proven and cost-effective approach is to discover SNPs by sequencing both DNA and RNA and subsequently genotype large numbers of individuals (Kumar et al., 2019; Lamichhaney et al., 2012; Montes et al., 2013, 2015; Therkildsen & Palumbi, 2017). For these methods, attention is explicitly restricted to *transcriptomic SNPs*: Those contained inside expressed genes due to their higher functional relevance, rather than intergenic and intronic regions. The combined approach of DNA and RNA sequences to SNP discovery has obtained the highest nonmodel SNP validation rates to date, without requiring a reference genome, and its success is largely due to the accurate detection of intron–exon boundaries (IEBs), which can confound genotyping primer design (Wang et al., 2008; see Figure 1). The IEB detection method developed by Conklin, Montes, Albaina, and Estonba (2013), for example, relies on

computing statistically significant positions in the transcript where many genomic reads start or end, indicating possible IEBs.

Traditional approaches to gene annotation in general, and IEB detection in particular, are based on the annotation of a genome assembly. For example, the NCBI Prokaryotic Genome Annotation Process (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/, last accessed 2020-03-04) relies on the prediction of transcribable regions based on alignments to known transcripts and proteins, and ab initio predictors of coding and noncoding genes. A more popular solution is to align either transcriptome or RNA-Seq reads with a splice-aware aligner such as GMAP (Wu & Watanabe, 2005), and extract from the results the IEB coordinates.

An alternative approach to finding IEBs can be based on creating a splice graph, a mathematical representation of a transcriptome where exons are represented by nodes, IEBs as edges, transcripts as paths, and genes as the different connected components. This approach is the one first presented by ChopStitch (Khan et al., 2018) where Bloom filters are used to store frequent *k*-mers of a shotgun whole-genome sequencing (WGS) dataset and use it to find signals of splicing in every sequence of a transcriptome assembly.

This paper presents EXFI, a memory-efficient tool for predicting and annotating the exons of a de novo transcriptome assembly through a splice graph representation. This tool works by comparing transcriptomic *k*-mers with those sequenced in a WGS experiment, marking potential IEBs wherever a section of a transcript is not found in it. To assess its performance, we compare it with ChopStitch and GMAP, using two synthetic datasets where references are available (human being and zebrafish); two fish species (Atlantic herring and Atlantic salmon) for which there exist reference annotations and experimental WGS datasets; and two species for which there only a draft genome and transcriptome are available (sugar pine and axolotl). Finally, we applied EXFI to a recently published dataset on tench (Kumar et al., 2019) to evaluate its success in IEB detection for SNP discovery in exonic regions.

We expect this method to be useful not only in the context of the original aim, the decomposition of transcripts into exons for gene-targeted SNP genotyping in organisms where genomic references are not available or not reliable, but also in the design of array-based tools such as sequence and exome capture, exome-wide genotyping, and RNA expression microarrays. Finally, recent developments in selective nanopore sequencing (Payne et al., 2020) are very likely to increase the relevance of exome-targeted approaches such as the one described here.

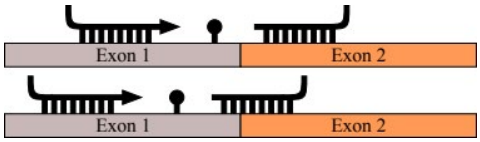


FIGURE 1 Two cases in primer design that can lead to genotyping failure: primers in different exons that require excessive PCR extension across an intron (top); a primer spanning an IEB will fail to anneal (bottom)

2 | METHODS

EXFI's core programs are written in Python, working on top of data processing (Pandas; McKinney, 2010) and Bioinformatics (BioPython; Cock et al., 2009) packages, as well as highly performant tools for *k*-mer manipulation (BioBloomTools' commit 0a42916, Chu et al., 2014; ABySS 2.0.2, Jackman et al., 2017; BEDTools 2.27.1, Quinlan & Hall, 2010). Its three main programs are build_baited_bloom_filter, build_splice_graph, and gfa1_to_fasta, to create the underlying data structure, to predict the splice graphs, and to write the exons, respectively.

2.1 | EXFI workflow

2.1.1 | Input

EXFI requires two input datasets: WGS reads and an assembled transcriptome in FASTQ and FASTA format, respectively. Such WGS reads may come from a single individual to multiple barcoded samples, even Pool-Seq approaches. The transcriptome assembly can be a published reference (Ensembl or NCBI Genomes, for example), or a de novo result from short-read or long-read sequencing technologies.

2.1.2 | Baited bloom filter construction

A Bloom filter (BF; Bloom, 1970) is a fast and succinct data structure for set membership (i.e., to test whether a *k*-mer is present in a transcript). Bloom filters have been successfully used in many high-throughput sequencing problems, including *k*-mer counting (Melsted & Pritchard, 2011), read compression (Benoit et al., 2015), read normalization (Crusoe et al., 2015), read filtering (Chu et al., 2014), error correction (Benoit, Lavenier, Lemaitre, & Rizk, 2014; Salmela & Rivals, 2014; Salmela, Walve, Rivals, & Ukkonen, 2017; Song, Florea, & Langmead, 2014), genome assembly (Chikhi, Limasset, & Medvedev, 2016; Chikhi & Rizk, 2012; Jackman et al., 2017; Peterlongo & Chikhi, 2012), gap filling (Paulino et al., 2015; Rizk, Gouin, Chikhi, & Lemaitre, 2014; Vandervalk et al., 2015), and targeted gene assembly (Kucuk et al., 2017). The advantage of this data structure is that it is very fast and space-efficient, with the drawback of being probabilistic: It does not return false negatives, but it can produce false positives with a tunable false-positive rate (BF FPR). This rate, for a given dataset, depends on three parameters that are under our control: the *k*-mer length, the amount of memory, and the number of hash functions used.

In the human and zebrafish genomes, only 4.24% and 5.68% of the bases are exons, respectively (Table 1). Therefore, this Bloom

filter approach can be used to remove WGS reads that are not exonic, and then reduce the BF FPR by nearly an order of magnitude. Additionally, cascading Bloom filters (Salikhov, Sacomoto, & Kucherov, 2014), a modification of original the data structure, stacks together multiple Bloom filters to keep frequent-enough *k*-mers and discard the ones produced by sequencing errors. Together, both approaches serve to filter out irrelevant but significant fractions of the original WGS experiment.

In EXFI, build_baited_bloom_filter uses both the transcriptome assembly and the WGS reads and performs the task in three steps. First, a Bloom filter of the transcriptome is built with biobloommaker. Second, each read of the WGS dataset that does not share at least one *k*-mer with the transcriptome is discarded with biobloomcategorizer. And third, the remaining reads are used to build a cascading Bloom filter with ABySS. The result is a binary file encoding the error-free *k*-mers of the reads that overlap the transcriptome.

2.1.3 | Exon and splice graph prediction

The exon and splice graph prediction procedure is carried out by the build_splice_graph script, which predicts in one step the exon sequences, the exon composition of each transcript, and the splice graph structure of the entire transcriptome.

First, transcriptomic *k*-mers are inspected sequentially: Those that overlap two different exons should not be present in the WGS dataset (Figure 2a) and therefore mark where an exon ends and the following starts. Then, consecutive positive *k*-mers that overlap by $k - 1$ bases are merged together, providing a draft exome (Figure 2b). The false positives that the Bloom filter produces may cause additional nucleotides in the raw exome and disconnected exons of length *k*. To prevent downstream problems, exons of length less than $k + q$ (q by default five) are filtered out (Figure 2c). Once deleted, a more relaxed merging step is applied when exons overlap by an excessive number of bases (10 by default; Figure 2d). Finally, if the -polish flag is specified, each pair of exons with a long-enough overlap is inspected for the donor/acceptor sites (usually GU/AG; 2e) and correctly trimmed if possible.

The primary output is a GFA1 file that encodes the inferred exons in terms of sequence and coordinates, the connections between them, and the transcripts as paths of exons. This type of file can be visualized with Bandage (Wick, Schultz, Zobel, & Holt, 2015), which also is helpful to manipulate exons and transcripts of interest, as well as to perform BLAST queries. Additionally, (gfa1_to_fasta) extracts the exons in FASTA format. It can also return the spliced transcripts, where each one of them is represented by the corresponding exons separated by a predefined amount of Ns.

2.2 | Validation datasets

Four reference datasets were selected: zebrafish (*Danio rerio*) and human being (*Homo sapiens*) as the key species, due to the depth of their available annotations; and Atlantic herring (*Clupea harengus*) and Atlantic salmon (*Salmo salar*), both with complete assemblies and exon annotations. Also, *Salmoniformes* are known to have an additional genome duplication round not shared by the other fish species here studied (Allendorf & Thorgaard, 1984), expanding both the genome length and number of genes (and therefore transcriptome complexity; Table 1). Additionally, to serve as a bridge between reference and de novo transcriptome assemblies, an RNA-Seq muscle library from Atlantic herring was assembled. Finally, two species without annotations, sugar pine (*Pinus lambertiana*) and axolotl (*Ambystoma mexicanum*), were added to test the upper limits of the methods studied in terms of genome length and sequencing effort. These two species are known for their large genome sizes (27 and 32 GB, respectively) due

TABLE 1 Experimental statistics of the studied cases

Experiment	Zebrafish	Human being	Atl. salmon	Atl. herring	Sugar pine	Axolotl	Tench
Genome type	Chromosome	Chromosome	Chromosome	Scaffold	Scaffold	Chromosome	Not available
Genome size (Gbp)	1.34	3.09	2.97	0.81	27.60	32.40	0.78
Genes	25,497	21,407	79,030	25,135	Unknown	Unknown	Unknown
Transcriptome type	Reference	Reference	Reference	Reference/de novo	De novo	De novo	De novo
Transcripts	51,714	164,776	109,584	29,353/97,777	331,11	180,605	267,058
Transcriptome size (Mbp)	110.69	270.48	355.21	64.18/55.39	36.74	229.48	294.70
Exons	495,200	1,199,596	1,313,909	314,220/Unknown	Unknown	Unknown	Unknown
Samples	2	6	20	50	1	1	10
Reads (M)	720.00	2,160.00	1,259.27	418.73	9,300.90	7,121.91	318.72
Total bases (Gbp)	72.00	216.00	125.93	41.13	1,395.13	712.19	31.87
Coverage	53.73	69.90	42.44	50.92	50.54	21.98	51.58

Note: Genome sizes are the number of characters in their corresponding reference files. All species are diploid.

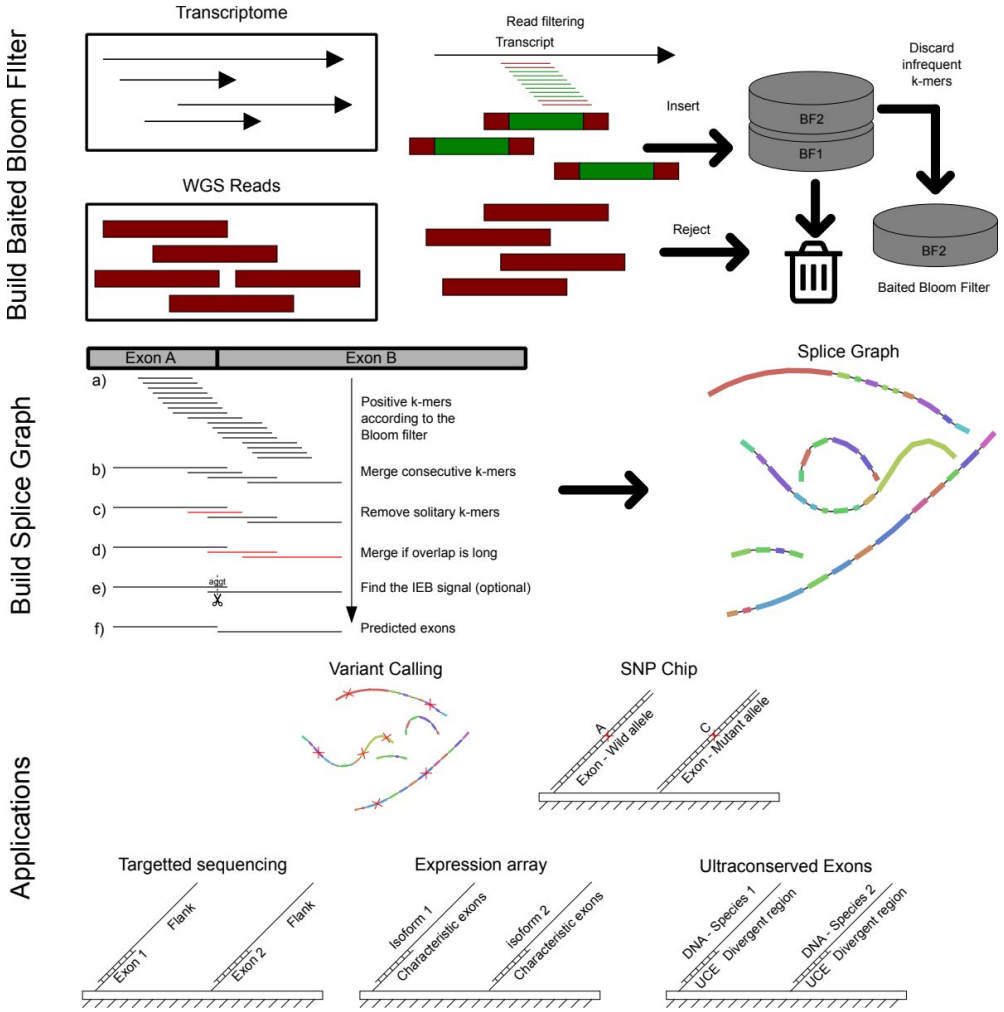


FIGURE 2 Schematic representation of the computational procedure. In the building stage, a WGS reads are filtered out according to whether or not they share a *k*-mer with the transcriptome. Positive reads are inserted into the cascading Bloom filter. Only the last one is used for analysis. The prediction step is comprised of multiples steps in which: (a) every transcriptomic *k*-mer is queried one by one to the filtered WGS set. (b) *k*-mers overlap by $k - 1$ bases are merged together. (c) Exons that are likely to be false positives are thrown away by a minimum length criterion. d) Exons that overlap by too many bases (ten by default) are merged together. (e) Overlaps between pairs of exons are inspected to see whether it contains the AG-GT splicing signal. (f) Exons are reported. Potential applications of EXFI include exome variant calling, design of SNP chips, targetted sequence of the exome, expression arrays, and UCE assays

to the extent of their repeat content (79% and 65.6% are transposable elements; Nowoshilow et al., 2018; Stevens et al., 2016).

Genomes, transcriptomes, and GFF3 annotations of *D. rerio* and *H. sapiens* were downloaded from Ensembl (release 91, assemblies GRCz10 and GRCh38, respectively; Kersey et al., 2018). Assembled genomes, transcriptomes, and annotations from *S. salar* (assembly GCA_000233375.4) and *C. harengus* (assembly GCA_000966335.1) were downloaded from NCBI Genome. Finally, in the case of *A. mexicanum* (assembly GCA_002915635.2) and *P. lambertiana* (GCA_001447015.2 assembly), assemblies were also downloaded from NCBI Genome, while their assembled transcriptomes were taken from the European Nucleotide Archive (accession numbers

GFZP01, Nowoshilow et al., 2018; and GEUZ01, Gonzalez-Ibeas et al., 2016, respectively).

With respect to humans and zebrafish, WGS reads were simulated with wgsim (Li, 2018), while for the other species, they were downloaded from different studies (Atlantic salmon: Kijas et al. (2018); Atlantic herring: Lamichhaney et al. (2012); Ambystoma mexicanum: Keinath et al. (2015); and Pinus lambertiana: Neale et al. (2014); full accession numbers available in Table S1). These assemblies varied in terms of both sequencing depth and individuals, from a 22× of a single individual in axolotl to 51× of a pool of 50 Atlantic herring samples.

2.3 | Benchmarking metrics

The performance of EXFI was compared with two tools: GMAP (Wu & Watanabe, 2005) and ChopStitch (Khan et al., 2018). GMAP is a method used to perform gapped alignments of expressed sequence tags (ESTs) and assembled transcripts to a reference genome. Its main advantage is that is easy to use and has been used extensively to annotate eukaryotic genomes with PASA (Haas et al., 2003). ChopStitch is a tool similar to EXFI that uses Bloom filters to predict exons and the splice graph, using the entire WGS dataset, and different exon prediction algorithms. Table 2 shows the main differences, advantages, and disadvantages between the three methods.

To compare the three methods in terms of speed and accuracy, two metrics were studied: one based on the recovery of the available annotation, and another in terms of mappability of the predicted exons to the genome.

For studying the recovery of the available annotation, reference exon coordinates in GFF format were transformed to BED, converting the chromosome-based coordinates to transcript-based, taking into account the strand and order of the exons. For example, if a pair of consecutive exons in transcript A in chromosome 1 are 1:1,000–1,100 and 1:1,500–1,600, they become A:0–100 and A:100–200 with respect to the transcriptome. Once converted, reference and predicted coordinates were compared with the BEDTools intersect subcommand, requiring a mutual overlap of at least 95% of the coordinates. With this program, the standard classification metrics are computed: precision (P = TP/(TP + FP), where TP and FP are the true and false positives, respectively), recall (R = TP/(TP + FN), where FN are the false negatives), and F_1 score (the harmonic mean between precision and recall: F_1 = 2PR/(P + R)). These comparisons are provided in the EXFI package via the compare_to_gff3 script.

TABLE 2 Qualitative comparison between the three methods studied: GMAP, ChopStitch, and EXFI

Software	GMAP	ChopStitch	EXFI
Input	Genome assembly (FASTA) Transcriptome (FASTA)	WGS reads (FASTQ) Transcriptome (FASTA)	WGS reads (FASTQ) Transcriptome (FASTA)
Output	Alignments (SAM, GFF3)	Exons (FASTA) Splice graph (DOT)	Splice graph (GFA) Exons (FASTA) Gapped transcripts (FASTA)
Steps	Genome index construction Spliced alignment Microexon identification	*k*-mer cardinality estimation Bloom filter construction Exon prediction Error correction Short exon prediction Splice graph construction	*k*-mer filtering Bloom filter construction Exon prediction Splice site polishing
Conda?	Yes	No, but via Brew	No, but via Dockerfile
Usability	Easy: index and predict	Easy: build and predict	Easy: build and predict
Genome input	Assembly	WGS	WGS
Sample variability	Genome and transcriptome may come from different sources	Transcriptome and WGS must come from the same individual	WGS reads can come from a Pool-Seq approach
Large genomes?	Yes (gmapl)	No	Yes
Speed	Fastest (minutes)	Medium (hours)	Slowest (hours)
Memory footprint	Medium/high	Medium/high	Low, adaptable
Precision/recall	Lowest	High	Highest
Mappability	High	High	Highest
Memory–FPR trade-off	—	Provide FPRs, then reserve optimal memory (may not be available)	Reserve memory, then return the FPR (may be too high).
Main advantage	Popular: easy to install	Fastest genome-free method	Memory and user-friendly, most accurate
Main disadvantage	Requires a genome assembly	Highest memory usage	Slowest of the methods

In the case of mappability measurements, predicted exons were aligned against their genomic reference with BWA MEM (Li, 2013), results were stored in BAM format with SAMTools (Li et al., 2009), and the reported statistics were obtained from the number of mapped exons, and the ones mapped with a perfect CIGAR string (all matches or with small insertions and deletions, but no base clipping).

2.4 | Objectives of the benchmarks

Before comparing the three methods, it was necessary to measure the influence of four parameters that may impact performance, in terms of both time and memory consumed, and the trade-off between precision and speed. From the two metrics described above, we used the annotation-based statistics as the ones that drove the experimental design since they showed more differences in terms of percentage points, and because mapping methods require a minimum seed length, which impacts the alignment of microexons.

First, the gains in terms of the BF FPR and exon prediction capabilities when reads are filtered or not were studied. Exons form a small fraction of a genome and only WGS reads that overlap the transcriptome are necessary to detect IEBs, while the remainder only increase the memory and BF FPR unnecessarily. The read filtering step implemented in EXFI retains not only exonic reads but also those in the flanking regions, where donor/acceptor signals and small variants can be detected. Therefore, EXFI was applied with and without the read filtering step, fixing the *k*-mer length to 25 bp, to measure (a) how much it accelerates or slows down the pipeline, (b) the BF FPRs, and (c) the fitness of the predicted exons.

Second, the effects of memory usage were compared. Next-generation sequencing projects are usually executed in high-performance computing environments, where RAM memory exceeds orders of magnitude what can be found in desktop and laptop computers. Probabilistic data structures such as Bloom filters have promised great savings in terms of memory, and therefore enabling analyses outside a computing cluster. To explore accuracy under different memory settings, EXFI was executed using the zebrafish dataset multiple times by varying the size of the Bloom filters from 4 to 60 GB in steps of 4 GB, and fixing with the *k*-mer length to 25 base pairs.

Third, the trade-off in terms of precision and recall with varying BF *k*-mer lengths was analyzed. If *k*is set too low, *k*-mers become less specific and more reads are inserted into the filter, increasing the BF FPR and lowering the precision, while increasing runtime too since there are more k-mers and reads processed. On the contrary, if *k* is set too high there will be fewer elements to insert, and since a significant fraction of them will contain variants and sequencing errors, they will be filtered by their low frequency (lowering the BF FPR but also the recall). To find the appropriate *k*-mer length, EXFI was run with the lowest and highest memory settings (4 and 60 GB) and by varying the *k*-mer length from 21 to 65 using odd values.

Finally, an acceptable genome coverage is needed for a successful experiment. On the one hand, a WGS experiment with little coverage will make the method underperform. On the other hand, too much coverage will make the BF FPR larger than necessary because of sequencing errors. As depth increases, the total number of true *k*-mers reaches a plateau, while the number of *k*-mers that contain sequencing errors keeps growing linearly (see figure 3 in Melsted & Pritchard, 2011). Therefore, a central point must exist in between to achieve near-optimal exon precision and recall values. The zebrafish datasets were sampled in 10% increments with Seqtk (Li, 2018), applying the procedure to each subsample, and measured the classification metrics, using both low and high memory settings and *k* fixed to 25 bp.

With respect to the other tools, GMAP version 2018.07.04 was executed using default parameters, and ChopStitch version 1.0.0, using the default *k*-mer length (50 bp) when possible, and varying the BF FPR values (and therefore different memory usages), over the six datasets (zebrafish, humans, and Atlantic herring), and we measured the performance in terms of the metrics described above: comparison against the annotations and mapping against the genome. All programs were run on a 2× Intel Xeon E5-2620 server, running in total 24 2 GHz threads, with 64 GB of RAM.

2.5 | Retrospective analysis of IEB prediction in *Tinca tinca*

To further validate EXFI for downstream analysis, the method was applied to retrieve the set of 96 transcriptomic SNPs in tench (*T. tinca*) wherein an earlier study (Kumar et al., 2019) was explored, where 92 of 96 were successfully genotyped. EXFI was executed using the assembled tench transcriptome, and the raw genomic reads comprised of two pools of five diploid individuals each, with an overall genome coverage of 52×. Finally, raw reads were mapped to the predicted exons with Bowtie2 (Langmead & Salzberg, 2012), and we performed SNP calling with BCFTools (Li et al., 2009). To derive the genotypable regions of the exons, variants with a quality value below 20 were filtered out, and then those that were within 35 bp to another variant or a predicted exon boundary.

3 | RESULTS

3.1 | Human and zebrafish simulations

As a practical approach, for each species a single Illumina HiSeq 2000 run per individual was simulated (360M PE reads), creating WGS datasets with coverages of 54× (2 runs, 720M PE reads) and 70× (6 runs, 2.21B PE reads; Table 1).

3.2 | Effects of read filtering

Filtering the reads resulted in a 68%–75% reduction of the BF FPR while also slightly improving all the classification metrics (Figure 3 and Table S2). We can observe a benefit of the filtering in the low memory case, where the FPR fell from 32.6% to 8.1% and rose the F_1 score from 89.8% to 94.8 when the maximum is of 95.6%. Additionally, we observe a slight reduction in time: from 172–186 to 149–173 min (Table S8). Therefore, filtering improves both the processing time and the prediction metrics. Similar conclusions can be reached in the human dataset (Table S3 and Figure S1).

3.3 | Effects of memory usage

The most significant parameter that impacts the Bloom filter is its size. Figure 3 and Table S2 show the expected decrease in BF FPR as space grew, but surprisingly, the exon precision and recall increased very slowly. Concretely, the BF FPR varied from 8.1% to 0.4% as the memory increased, achieving a 95.8% precision and 93.8% recall in the low memory case, when in the high memory case one both values were respectively 96.6% and 94.6% (Table S2). With respect to the human dataset, experiments were only performed with the low and high memory settings, obtaining BF FPRs of 13.7% and 0.7%, achieving 93.1% and 94.7% precision, and 89.3% and 90.9% recall, respectively (Table S3 and Figure S1). Therefore, a 4 GB Bloom filter is enough to achieve near-optimal results. Also, it is not necessary to demand particularly low (5% or less) BF FPR to predict exons accurately.

3.4 | Effects of *k*-mer length

As described in Methods sections, precision and recall are related through the *k*-mer length. On the one hand, as *k* increases, the precision also increases until $k = 47$, when it starts to decrease rapidly (Figure 4 and Table S4). On the other hand, recall decreased almost from the start (4 GB: $k = 25$; 60 GB: $k = 23$). According to the F_1 statistic (the harmonic mean between precision and recall), for both methods, there is a window of *k*-mers, from 23 to 35, where this metric remains stable, boosting the recall when *k* is small, and the precision when k is high. Given the results, we used for the remainder of the analysis a *k*-mer length of 25 bp to keep the recall as high as possible while keeping precision high too, and set it as the default value in EXFI.

3.5 | Effects of sequencing depth

The sequencing depth increased the power of the precision up to a certain point (Figure 5 and Table S5). For a 16× sequencing depth, precision and recall already are above 90%, and maximum values

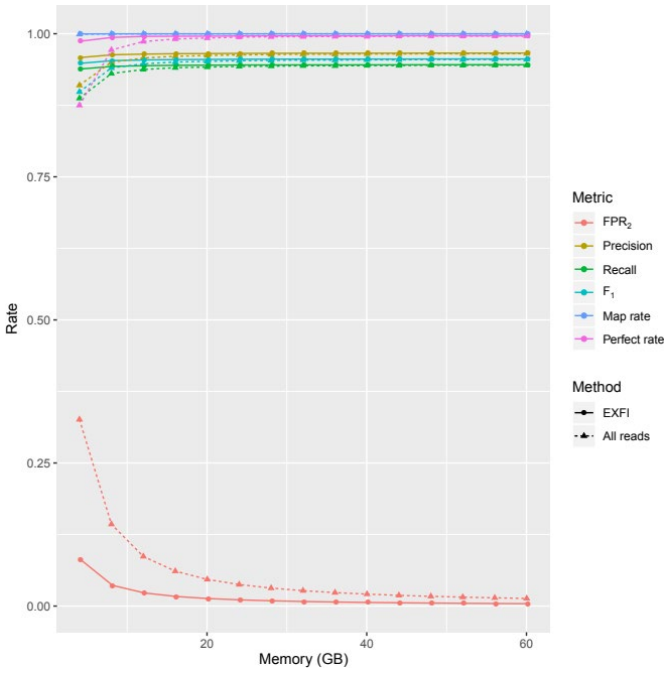


FIGURE 3 Classification and mapping rates of EXFI depending on Bloom filter sizes in the zebrafish dataset. Filtering the dataset yields better classification and mapping rates by lowering the FPR_2. These values are already near-optimal when four Gigabytes are allocated. The raw mapping rates were close to 100% from the start for both methodologies. For the perfect mapping rates, we see EXFI achieving a 98.75% mapping rate from the start

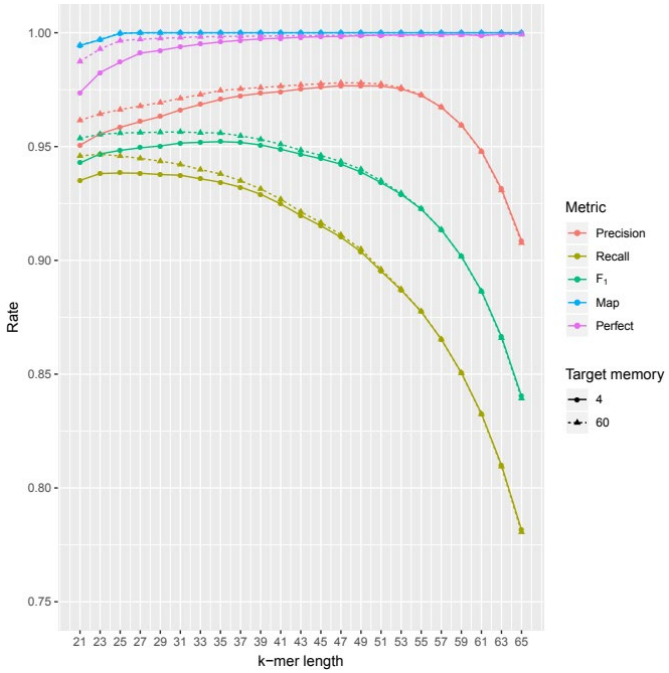


FIGURE 4 Precision and recall of EXFI when the *k*-mer length varies, using the minimum and maximum memory settings. As expected, the longer the *k*, the higher the precision, and the lower the recall. For both methods, best results are achieved when the *k*-mer length varies between 23 and 35. Perfect mapping rates ranged from 97.37% to 99.95%

(96.7% and 94.7%, respectively) are reached when coverage is between 26× and 37×. Past that coverage window, precision and recall start decreasing due to sequencing errors, and unnecessarily raising the BF FPR. Therefore, a sequencing depth of at least 20× is good enough, and that optimally should be between 30× and 40× to retrieve exons from a transcriptome with EXFI.

3.6 | Comparison with ChopStitch and GMAP on simulated and real datasets

The performance of ChopStitch, EXFI, and GMAP was compared across six species in terms of the BF FPR and sizes, classification, and mappability scores. Given the results above, we chose to run EXFI using 4 GB of RAM, and a *k*-mer length of 25. For ChopStitch, we used the default *k*-mer length of 50 bp, and default BF FPRs of 1% when possible. For GMAP, the default parameters were the ones used. In the case of the megagenomes, gmapl was used as the alignment tool.

There are several differences between EXFI and ChopStitch. Algorithmically, in EXFI the total amount of memory to be used is specified at the beginning, the number of hash functions is fixed (to four, fixed number in the version of ABySS used), the reads are filtered and processed, and the BF FPR is returned at the end. In contrast, the reverse procedure is applied in ChopStitch: The desired BF FPRs are first specified, and the optimal sizes and number of hash functions are estimated from the full dataset of reads. This procedure selects the optimal memory (maybe unavailable) and number of hash functions to work, but requires to process twice the full WGS reads: one for estimation and other for actual computations. On the other side, EXFI hashes all the WGS reads in two steps: once for the filtering purpose, and a second time for the remainder.

In zebrafish, we considered running EXFI and ChopStitch with multiple memory/BF FPR configurations (respectively 4–60 GB in 4 GB increments, and FPR_1 varying from 20% to 1% and BF FPR_2 set to 1%). In general, EXFI outperformed both methods (Figure 6 and Table S6) and its performance remained high and constant from 4 GB.

When comparing EXFI's low memory mode against ChopStitch default 1% FPR_2 (and 28 GB) and GMAP (Table 3), we observe that with a BF FPR_2 of 8% (and 4 GB), EXFI obtained a slightly better F_1 score (with better precision and worse recall) than the other two methods. According to the exon mappability, more than 98% of the predictions of both reference-free methods were perfectly matched to the genome, while the reference-based tool obtained 92.7%.

With respect to the human dataset, all methods obtained lower metrics than in the zebrafish case, due both to the higher complexity of the transcriptome and the length of the genome. With the default settings, EXFI outperformed both methods with an exon F_1 score of 91.2%. Due to the number of different *k*-mers to process, ChopStitch's default *k*-mer length value had to be lowered to 25 and the target BF FPR_1 had to be raised to 15% in order to avoid memory

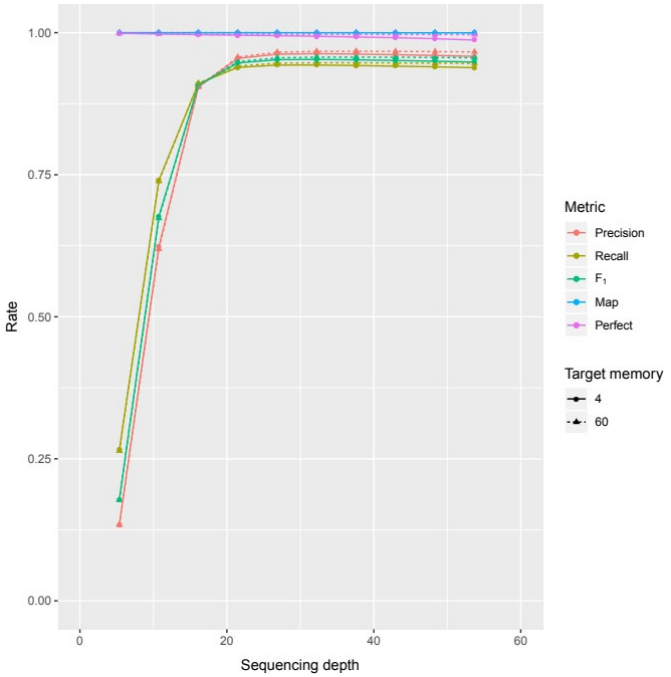


FIGURE 5 Precision and recall values of EXFI depending on the sequencing depth, using the minimum and maximum memory settings and the *k*-mer length fixed to 25. Both settings produced similar results, obtaining higher metrics when all the memory was used. Around 25–30× almost all error-free *k*-mers are sampled, and then, sequencing errors start to pollute the Bloom filter. Both mapping rates stayed above 98.7%

allocation errors. In this case, ChopStitch obtained an F_1 of 88.6%, and GMAP of 88.3% (Table 3, Table S7, and Figure S2).

In both datasets, GMAP obtained the fastest data structure construction (24 and 53 min to index the zebrafish and human genomes; Table S8), followed by ChopStitch (2 hr 38 min and 4 hr 22 min) and EXFI (2 hr 29 min and 6 hr 18 min). On the other hand, GMAP finished last when predicting (more than 15 min in both cases) and using 24 threads, while for a single compute thread, ChopStitch was the fastest (3 min 41 s and 7 min 18 s in zebrafish and human cases) followed by EXFI (5 min 56 s and 14 min 30 s).

Similar results were obtained when analyzing the salmon transcriptome (Table 3): EXFI obtained the lightest RAM consumption with the cost of obtaining a higher BF FPR_2 (4.23%), while ChopStitch achieved a 1% BF FPR_2 with 8.7 GB of RAM. With respect to the prediction of exons, EXFI obtained higher classification and mapping scores, followed by ChopStitch and GMAP.

In the Atlantic herring reference dataset, we observe that both *k*-mer-based methods obtained worse-than-expected F_1 scores than GMAP when analyzing the reference transcriptome, while still obtaining the highest perfect mappings (in EXFI's case, the highest across all datasets, 99.5%). In the de novo transcriptome case, predictions of all three methods had lower mapping rates than the reference case, with ChopStitch leading the comparison with 87.1% of perfectly mapped transcripts, followed by EXFI (82.3%) and GMAP (57.8%).

Finally, for the axolotl and sugar pine megagenomes, we did not obtain results for all of the methods. Due to the terabase pairs sequenced and the size of the references, ChopStitch was only able to produce a Bloom filter for the axolotl, with a BF FPR_2 of 20.2% and a *k*-mer length of 25 bp, and GMAP was able to build both references but failed to produce predictions in the axolotl case due to memory exhaustion. For EXFI, even though it can produce Bloom filters with 4 GB of RAM, the BF FPR_2s were too high to work (52% and 29.4%, respectively; data not shown), and therefore, we raised the RAM to 60 GB to obtain reasonable FPRs. Indeed, we obtained data structures with BF FPRs of 3.1% and 2.0% in the Pine and axolotl cases and after 2 and 1 days of execution, respectively. In the sugar pine case, 90.3% of the exons were perfectly mapped to the reference (99.7% when clippings were allowed), while GMAP obtained lower results (95.7% mappable, 67.3% without trimming alignments). With respect to axolotl, 78.2% of EXFI's predictions were matched end to end to the genome (and 98.8% at least in part), while ChopStitch obtained a 77.2% rate (85.1% when clipping was allowed).

3.7 | Retrospective analysis in *T. tinca*

From the set of 266,578 input transcripts, EXFI predicted 1,072,772 exons. In total, after quality and distance filtering, 228,931 SNPs and 26,169 indels were predicted suitable for genotyping. All IEBs

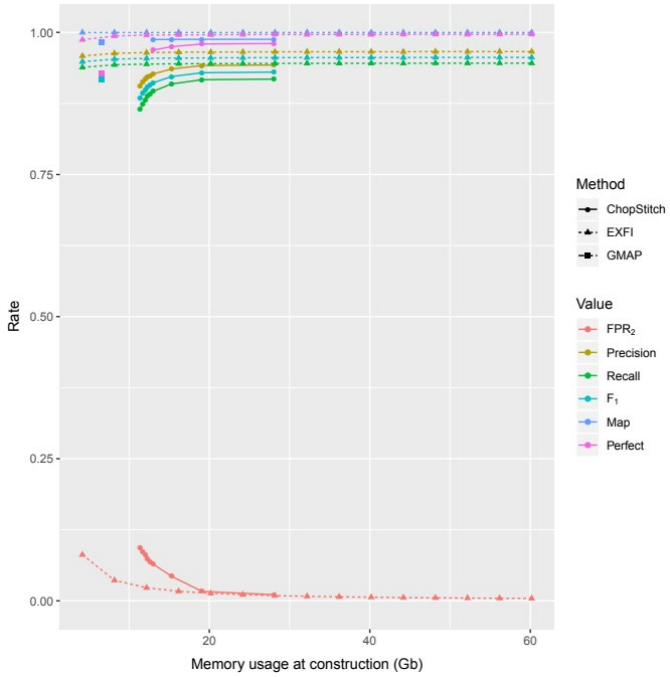


FIGURE 6 Comparative on memory and precision/recall metrics between ChopStitch, EXFI, and GMAP. EXFI's performance remained high and constant from the start

proximal to the 96 SNPs described in Kumar et al. (2019) using the Conklin et al. (2013) method were detected by EXFI (therefore 100% precision over this set of SNPs). One SNP was proximal to a false-positive EXFI IEB, due to multiple variants in a short space, indicating that it would not have been selected for genotyping primer design. Therefore, on this retrospective SNP discovery task, EXFI would recall 95 of 96 of the selected SNPs.

4 | DISCUSSION

We developed EXFI, a method that reliably predicts the exon sequences and splice graph of a species using a de novo-assembled transcriptome and raw WGS reads. We tested it in multiple eukaryotic species, varying the genome and transcriptome reference status, simulated and experimental datasets, and samples with different level of heterozygosity of the samples. We found out that EXFI performs better in terms of memory and classification than other tools when describing the structural annotation of every transcript.

We studied the four principal parameters that can affect the prediction procedure: read filtering, memory, *k*-mer length, and genome coverage. First, by filtering the transcriptome, we ended up reducing by two-thirds the BF FPR while also slightly decreasing the execution time. Therefore, this reduction can be translated into a memory optimization. Second, using more than 4 GB of RAM (and higher BF FPR) yielded equally accurate predictions as using 60 GB (Figures 3 and 6). Thus, commodity desktop and laptop computers are enough to achieve accurate exon predictions on gigabase-sized genomes. Third, our approximation predicted a window of optimal *k*-mer length values between 23 and 35 base pairs. Finally, we show that 20× coverage is good enough for exon prediction, with optimal coverage between 30× and 40×.

We compared EXFI against ChopStitch, a similar method, and GMAP, a splice-aware program designed to align transcripts to a reference genome. We used datasets that vary in genome size, sequencing depth, number of individuals, and type of input transcriptome. When taking into account the general picture across the different reference species (zebrafish, humans, salmon, and herring), even with higher BF FPR rates, EXFI obtained better prediction metrics, except for the Atlantic herring (Table 3). In that case, EXFI was less accurate predicting exons, but in terms of mappability, it achieved the highest results across all datasets (99.5%). This high mappability result, although more optimistic than the exon precision, means that even if the exon prediction is not precise, it is still usable for downstream analysis. These perfectly matched predictions are interior sections of the exons rather than the full sequence, which makes them suitable for genotyping, array design, and sequence capture.

We also studied three situations where the input transcriptome was de novo-assembled from RNA-Seq reads: Atlantic herring, to compare the differences between reference and assembled transcriptomes, and the megagenomes of axolotl and sugar pine. In terms of mappability, all three methods performed worse than in reference

TABLE 3 Performance of the three tools across different species

Species	Method	Time	Memory	FPR1	FPR2	Precision	Recall	F_1	Mapped	Perfect
Zebrafish	ChopStitch	1 hr 41 min 54 s	28.060	0.010	0.011	0.943	0.918	0.930	0.988	0.980
Zebrafish	EXFI	2 hr 35 min 15 s	**4.177**	0.256	0.081	**0.958**	**0.938**	**0.948**	**1.000**	**0.987**
Zebrafish	GMAP	**40 min 19 s**	6.567	–	–	0.918	0.917	0.917	0.982	0.927
Human being	ChopStitch*	4 hr 28 min 58 s	30.424	0.158	0.100	0.903	0.868	0.886	0.988	**0.969**
Human being	EXFI	6 hr 32 min 49 s	**4.364**	0.361	0.137	**0.931**	**0.893**	**0.912**	**1.000**	0.957
Human being	GMAP	**1 hr 11 min 25 s**	9.301	–	–	0.883	0.884	0.883	0.985	0.907
Herring R.	ChopStitch	49 min 53 s	5.679	0.010	0.011	0.819	0.858	0.838	0.974	0.965
Herring R.	EXFI	1 hr 25 min 6 s	**4.123**	0.064	0.024	0.816	0.866	0.840	**1.000**	**0.995**
Herring R.	GMAP	**19 min 8 s**	4.707	–	–	**0.949**	**0.941**	**0.945**	0.983	0.935
Herring A.	ChopStitch	50 min 2 s	5.705	0.010	0.011	–	–	–	0.972	**0.871**
Herring A.	EXFI	1 hr 32 min 8 s	**4.111**	0.068	0.026	–	–	–	**0.986**	0.823
Herring A.	GMAP	**37 min 20 s**	6.564	–	–	–	–	–	0.921	0.578
Salmon	ChopStitch	2 hr 57 min 38 s	8.657	0.010	0.010	0.883	0.887	0.885	0.985	0.975
Salmon	EXFI	4 hr 49 min 37 s	**4.466**	0.080	0.042	**0.901**	**0.904**	**0.903**	**0.999**	**0.987**
Salmon	GMAP	**1 hr 22 min 15 s**	9.320	–	–	0.809	0.830	0.819	0.979	0.866
Sugar pine	ChopStitch*	–	–	–	–	–	–	–	–	–
Sugar pine	EXFI	2 days 7 hr 38 min 57 s	60.090	0.090	0.031	–	–	–	**0.997**	**0.903**
Sugar pine	GMAP	**6 hr 20 min 13 s**	**55.371**	–	–	–	–	–	0.956	0.673
Axolotl	ChopStitch*	**14 hr 29 min 38 s**	**29.629**	0.202	0.142	–	–	–	0.851	0.772
Axolotl	EXFI	1 day 3 hr 20 min 50 s	60.313	0.040	0.020	–	–	–	**0.988**	**0.782**
Axolotl	GMAP	–	–	–	–	–	–	–	–	–

Note: Best metrics across the three methods are marked in bold. Time is the sum of the walltimes at the building and prediction steps. When possible, the steps were run using all processors available, that is, in ChopStitch's and EXFI's build steps, and in GMAP's predict stage. Memory, expressed in Gigabytes, represents the peak usage in memory. FPR represents the false-positive rate of the Bloom filter used for prediction. Mapped and Perfect stands for the overall alignment rate of the predicted exons, allowing and not allowing clipping, respectively. EXFI was executed to use only 4 GB of RAM except for the megagenomes. ChopStitch with *k*-mer lengths of 50 bp and FPRs of 1%, except when memory usage was an issue. In the cases marked with asterisks, the *k*-mer lengths were lowered to 25 bp, and target FPR values were tested one by one in the set of 1%, 5%, 10%, 15%, and 20%. Actual FPRs are the ones reported. In general, when a reference transcriptome was used, EXFI obtained the best precision, while ChopStitch obtained better recall. With respect to alignments to the genomes, EXFI obtained the best mapping rates.

cases due to the inherent complexity of transcriptome assembly. In the herring case, the mappability scores fell for all methods. In the axolotl case, we obtained moderate results for EXFI (78.2% perfect mapping) and ChopStitch (71.4%). Finally, in the sugar pine dataset, EXFI's performance stood high (90.3%), while GMAP did moderately (67.3%). Interestingly, results in herring and salmon suggest that reference-free method remain accurate even when WGS datasets come from a Pool-Seq approach. Another lesson learned is that special care has to be taken regarding the input transcriptome. While the axolotl and sugar pine transcriptomes come from a wide variety of tissues and conditions and are sequenced in-depth, the herring transcriptome was obtained from a single tissue, where its characteristic transcripts were assembled in full length, but where the lowly expressed ones appear fragmented, and specific transcripts to other tissues are missing.

Finally, this paper also studied the performance of EXFI in an earlier transcriptomic SNP discovery project in tench (Kumar

et al., 2019). EXFI was also able to find hundreds of thousands of SNPs across almost a million exons. With respect to the set of known genotyped exons, EXFI obtained 100% precision and 99% recall.

These positive results for EXFI are due to the read filtering step and the exon prediction rules used. The filtering step is critical in eukaryotic genomes because a significant fraction of the WGS dataset is not only unnecessary but misleading. The convenience of the exon prediction rules is extracted from Figures 3 and 6: When comparing ChopStitch and EXFI without read filtering, the latter obtained slightly superior precision and recall due to the exon prediction methods, in spite of the relatively high BF FPR (8% vs. 1%). Moreover, EXFI's predictions are accurate enough when working with relatively high BF FPR_2.

Previous structural annotation algorithms rely on a whole-genome assembly followed by the mapping of RNA-Seq reads, ESTs and transcripts, and homology predictions against genome, transcriptome, and protein databases. Our results suggest that EXFI is a reliable tool too while avoiding completely the step of generating a high-quality genome assembly.

Recent reviews have been published on RAD-Seq and Targeted Sequencing approaches (Harvey, Smith, Glenn, Faircloth, & Brumfield, 2016; Lowry et al., 2017; Meek & Larson, 2019) explaining the advantages and disadvantages of all methods, with the same conclusion: Targeted approaches should be preferred for large quantities of samples and loci. These methods have in common the enrichment of ultraconserved elements (UCEs; Faircloth et al., 2012) or exons under varying selection types. EXFI can be used for both approaches: Conservation of exons can be measured by orthology analysis against other exon predictions and known reference genomes, transcriptomes, and proteomes; and the different selective pressures can be obtained by performing variant calling on the exome given the set of WGS reads used in the analysis.

For optimal results, we propose a two-step experimental approach to study nonmodel exomes: an initial exploration of the exome structure and the variants it contains, followed by targeted sequencing of hundreds to thousands of samples. For the first step, it would be necessary to sequence RNA from as many tissues and development stages, aiming to get the best representation of the transcriptome, and to sequence between 30× and 40× of the genome, preferably from multiple individuals, to discover as many variants as possible. In this regard, Therkildsen and Palumbi (2017) have shown that is possible to move from pools of DNA to individually barcoded individuals. In a second step, a targeted approach would be obtained for thousands of loci and samples, leaving behind most of the genome and therefore being able to fit more individuals and populations in the same sequencing assay. As it happened for Atlantic herring, a DNA sequencing effort initially focused on the transcriptome (Lamichhaney et al., 2012) was reused years later once genome assembly was possible (Barrio et al., 2016).

This report has presented EXFI, a pipeline that predicts the splice graph and exon sequences from a transcriptome and WGS reads instead of a reference genome. Different parameters that affect its performance were studied: read filtering, memory usage, *k*-mer length, and sequencing depth. Tests were carried out on zebrafish and human simulations, Pool-Seq samples of Atlantic salmon and Atlantic herring, and the megagenomes of the sugar pine and axolotl, varying all in sequencing depth, heterozygosity, genome length, and complexity. A retrospective analysis of a recently published set of transcriptomic SNPs on tench was also done, obtaining 100% precision and 99% recall. It is shown that it is possible to perform structural annotation of a transcriptome of heterogeneous samples with low computational resources. Finally, EXFI is expected to be particularly useful for population genetic studies, phylogenetic relationships, and RNA expression in nonmodel species.

ACKNOWLEDGMENTS

We acknowledge support with the predoctoral grant (PRE_2017_2_0169) and the Genomic Resources Research Group (grant IT558-10), both funded by the Basque Government.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Jorge Langa: Conceptualization (equal); software (lead); validation (lead); visualization (lead); writing – original draft (equal); writing – review and editing (equal). **Andone Estonba:** Funding acquisition (lead); supervision (equal); writing – original draft (equal); writing – review and editing (equal). **Darrell Conklin:** Conceptualization (equal); funding acquisition (equal); supervision (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal).

DATA AVAILABILITY STATEMENT

EXFI is open source and freely available at https://github.com/jlanga/exfi. It is subject to Continuous Integration and Unit Testing. Detailed installation and usage are available in the README file of the repository (https://github.com/jlanga/exfi/README.md). Additionally, test data are included in the repository. Moreover, a Dockerfile is available at https://github.com/jlanga/exfi-docker to create a container with all the tools installed. Finally, the scripts to validate, benchmark, and reproduce the tables and figures in this document can be found online as a Snakemake pipeline (Köster & Rahmann, 2012) at https://github.com/jlanga/smsk_exfi_paper. Archived versions of the resources here described are available at Dryad (https://doi.org/10.5061/dryad.tx95x69vc).

ORCID

Jorge Langa https://orcid.org/0000-0001-5137-8204

REFERENCES

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389–400. https://doi.org/10.1534/genetics.115.183665

Allendorf, F. W., & Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In B. J. Turner (Ed.), *Evolutionary genetics of fishes*.

Monographs in evolutionary biology (pp. 1–53). Boston, MA: Springer, US. https://doi.org/10.1007/978-1-4684-4652-4_1

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Barrio, A. M., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife*, *5*, e12081. https://doi.org/10.7554/eLife.12081

Benoit, G., Lavenier, D., Lemaitre, C., & Rizk, G. (2014). Bloocoo, a memory efficient read corrector. In *European conference on computational biology (ECCB)*. Retrieved from https://hal.inria.fr/hal-01092960

Benoit, G., Lemaitre, C., Lavenier, D., Drezen, E., Dayris, T., Uricaru, R., & Rizk, G. (2015). Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph. *BMC Bioinformatics*, *16*(1), 288. https://doi.org/10.1186/s12859-015-0709-7

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, *13*(7), 422–426. https://doi.org/10.1145/362686.362692

Chikhi, R., Limasset, A., & Medvedev, P. (2016). Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, *32*(12), i201–i208. https://doi.org/10.1093/bioinformatics/btw279

Chikhi, R., & Rizk, G. (2012). Space-efficient and exact de Bruijn graph representation based on a bloom filter. In *Lecture Notes in Computer Science*, WABI, 7534 (pp. 236–248). Springer.

Chu, J., Sadeghi, S., Raymond, A., Jackman, S. D., Nip, K. M., Mar, R., … Birol, I. (2014). BioBloom tools: Fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics*, *30*(23), 3402–3404. https://doi.org/10.1093/bioinformatics/btu558

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Conklin, D., Montes, I., Albaina, A., & Estonba, A. (2013). Improved conversion rates for SNP genotyping of nonmodel organisms. In *International work-conference on bioinformatics and biomedical engineering (Iwbbio)* (pp. 127–134). Granada, Spain.

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., … Brown, C. T. (2015). The khmer software package: Enabling efficient nucleotide sequence analysis. *F1000Research*, *4*, 900. https://doi.org/10.12688/f1000research.6924.1

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726. https://doi.org/10.1093/sysbio/sys004

Gonzalez-Ibeas, D., Martinez-Garcia, P. J., Famula, R. A., Delfino-Mix, A., Stevens, K. A., Loopstra, C. A., … Wegrzyn, J. L. (2016). Assessing the gene content of the megagenome: Sugar pine (*Pinus lambertiana*). *G3: Genes*, *Genomes*, *Genetics*, *6*(12), 3787–3802. https://doi.org/10.1534/g3.116.032805

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr, L. I., Hannick, R. M. et al (2003). Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, *65*(5), 910–924. https://doi.org/10.1093/sysbio/syw036

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., … Birol, I. (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777. https://doi.org/10.1101/gr.214346.116

Keinath, M. C., Timoshevskiy, V. A., Timoshevskaya, N. Y., Tsonis, P. A., Randal Voss, S., & Smith, J. J. (2015). Initial characterization of the large genome of the salamander *Ambystoma Mexicanum* using shotgun and laser capture chromosome sequencing. *Scientific Reports*, *5*(November), 16413. https://doi.org/10.1038/srep16413

Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., … Yates, A. (2018). Ensembl genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*, *46*(D1), D802–D808. https://doi.org/10.1093/nar/gkx1011

Khan, H., Mohamadi, H., Vandervalk, B. P., Warren, R. L., Chu, J., & Birol, I. (2018). ChopStitch: Exon annotation and splice graph construction using transcriptome assembly and whole genome sequencing data. *Bioinformatics*, *34*(10), 1697–1704. https://doi.org/10.1093/bioinformatics/btx839

Kijas, J., McWilliam, S., Sanchez, M. N., Kube, P., King, H., Evans, B., … Verbyla, K. (2018). Evolution of sex determination loci in Atlantic salmon. *Scientific Reports*, *8*(1), 5664. https://doi.org/10.1038/s41598-018-23984-1

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kucuk, E., Chu, J., Vandervalk, B. P., Austin Hammond, S., Warren, R. L., & Birol, I. (2017). Kollector: Transcript-informed, targeted de novo assembly of gene loci. *Bioinformatics*, *33*(12), 1782–1788. https://doi.org/10.1093/bioinformatics/btx078

Kumar, G., Langa, J., Montes, I., Conklin, D., Kocour, M., Kohlmann, K., & Estonba, A. (2019). A novel transcriptome-derived SNPs array for tench (*Tinca tinca* L.). *PLoS One*, *14*(3), e0213992. https://doi.org/10.1371/journal.pone.0213992

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic Herring. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

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

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv:1303.3997 [Q-Bio], March. Retrieved from http://arxiv.org/abs/1303.3997

Li, H. (2018). wgsim - Read simulator for next generation sequencing. *Github repository*. http://github.com/lh3/wgsim

Li, H. (2018). Seqtk: Toolkit for processing sequences in FASTA/Q formats. *GitHub repository*.

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. https://doi.org/10.1093/bioinformatics/btp352

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152. https://doi.org/10.1111/1755-0998.12635

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (pp. 51–56).

Meek, M. H., & Larson, W. A. (2019). The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Molecular Ecology Resources*, *19*(4), 795–803. https://doi.org/10.1111/1755-0998.12998

Melsted, P., & Pritchard, J. K. (2011). Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics*, *12*, 333. https://doi.org/10.1186/1471-2105-12-333

Montes, I., Conklin, D., Albaina, A., Creer, S., Carvalho, G. R., Santos, M., & Estonba, A. (2013). SNP discovery in European anchovy (*Engraulis encrasicolus*, L) by high-throughput transcriptome and genome sequencing. *PLoS One*, *8*(8), e70051. https://doi.org/10.1371/journal.pone.0070051

Montes, I., Langa, J., Vilas, C., Helyar, S., Alvarez, P., Conklin, D., & Estonba, A. (2015). Discovery and characterization of 80 SNPs and 1,624 SSRs in the transcriptome of Atlantic mackerel (*Scomber scombrus*, L). *Molecular Ecology Resources*, *15*(6), 1510–1512.

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., … Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*(3), R59. https://doi.org/10.1186/gb-2014-15-3-r59

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., Pippel, M., … Myers, E. W. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature*, *554*(7690), 50–55. https://doi.org/10.1038/nature25458

Paulino, D., Warren, R. L., Vandervalk, B. P., Raymond, A., Jackman, S. D., & Birol, I. (2015). Sealer: A scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics*, *16*(1), 230. https://doi.org/10.1186/s12859-015-0663-4

Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B., & Loose, M. (2020). Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels. *bioRxiv*, February 2020.02.03.926956. https://doi.org/10.1101/2020.02.03.926956

Peterlongo, P., & Chikhi, R. (2012). Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. *BMC Bioinformatics*, *13*(1), 48. https://doi.org/10.1186/1471-2105-13-48

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

Rizk, G., Gouin, A., Chikhi, R., & Lemaitre, C. (2014). MindTheGap: Integrated detection and assembly of short and long insertions. *Bioinformatics*, *30*(24), 3451–3457. https://doi.org/10.1093/bioinformatics/btu545

Salikhov, K., Sacomoto, G., & Kucherov, G. (2014). Using cascading bloom filters to improve the memory usage for de Brujin graphs. *Algorithms for Molecular Biology*, *9*(1), 2. https://doi.org/10.1186/1748-7188-9-2

Salmela, L., & Rivals, E. (2014). LoRDEC: Accurate and efficient long read error correction. *Bioinformatics*, *30*(24), 3506–3514. https://doi.org/10.1093/bioinformatics/btu538

Salmela, L., Walve, R., Rivals, E., & Ukkonen, E. (2017). Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, *33*(6), 799–806. https://doi.org/10.1093/bioinformatics/btw321

Song, L. I., Florea, L., & Langmead, B. (2014). Lighter: Fast and memory-efficient sequencing error correction without counting. *Genome Biology*, *15*, 509. https://doi.org/10.1186/s13059-014-0509-9

Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., … Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics*, *204*(4), 1613–1626. https://doi.org/10.1534/genetics.116.193227

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194–208. https://doi.org/10.1111/1755-0998.12593

Vandervalk, B. P., Yang, C., Xue, Z., Raghavan, K., Chu, J., Mohamadi, H., … Birol, I. (2015). Konnector V2.0: Pseudo-long reads from paired-end sequencing data. *BMC Medical Genomics*, *8*(3), S1. https://doi.org/10.1186/1755-8794-8-S3-S1

Wang, S., Sha, Z., Sonstegard, T. S., Liu, H., Peng, X. U., Somridhivej, B., … Liu, Z. (2008). Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics*, *9*(1), 450. https://doi.org/10.1186/1471-2164-9-450

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de Novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Langa J, Estonba A, Conklin D. EXFI: Exon and splice graph prediction without a reference genome. *Ecol Evol*. 2020;10:8880–8893. https://doi.org/10.1002/ece3.6587

Appendix C

Article 3. European sardine

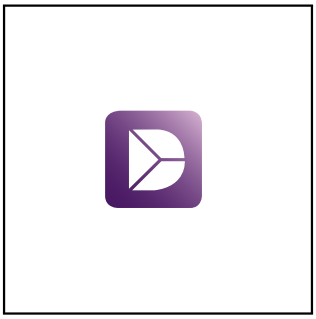
Langa, J., Huret, M., Montes, I., Conklin, D., & Estonba, A. (2021). Transcriptomic dataset for Sardina pilchardus: Assembly, annotation, and expression of nine tissues. *Data in Brief*, 107583. https://doi.org/10.1016/j.dib.2021.107583.

TABLE C.1: Quality Metrics for Data in Brief in 2020. Data in Brief is not indexed in the Web of Science database.

Data in Brief 2020 - Web of Science	
Category	-
Impact Factor	-
Rank	-
Quantile	-

Data in Brief 2020 - Scopus	
Category	Multidisciplinary
CiteScore	1.7
Rank	32/110
Quantile	Q2

Journal Pre-proof



Transcriptomic dataset for Sardina pilchardus: assembly, annotation, and expression of nine tissues

Jorge Langa , Martin Huret , Iratxe Montes , Darrell Conklin , Andone Estonba

PII: S2352-3409(21)00858-1
DOI: https://doi.org/10.1016/j.dib.2021.107583
Reference: DIB 107583

To appear in: *Data in Brief*

Received date: 2 June 2021
Revised date: 27 August 2021
Accepted date: 9 November 2021

Please cite this article as: Jorge Langa , Martin Huret , Iratxe Montes , Darrell Conklin , Andone Estonba , Transcriptomic dataset for Sardina pilchardus: assembly, annotation, and expression of nine tissues, *Data in Brief* (2021), doi: https://doi.org/10.1016/j.dib.2021.107583

Transcriptomic dataset for Sardina pilchardus: assembly, annotation, and expression of nine tissues

Authors

Jorge Langa[1], Martin Huret[2], Iratxe Montes[1], Darrell Conklin[3,4], Andone Estonba[1]

Affiliations

1. Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country, 48940 Leioa, Spain
2. IFREMER, STH/LBH, B.P. 70, Plouzané, 29280 France
3. Department of Computer Science and Artificial Intelligence, Faculty of Computer Science, University of the Basque Country UPV/EHU, San Sebastián, Spain
4. IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Corresponding author

Andone Estonba (andone.estonba@ehu.eus), Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country, UPV/EHU. 48940 Leioa (Bizkaia), Spain

Abstract

European sardine or pilchard is a planktonic small pelagic fish present from the North Sea in Europe to the coast of Senegal in the North of Africa, and across the Mediterranean sea to the Black Sea. Ecologically, sardines are an intermediary link in the trophic network, preying on plankton and being predated by larger fishes, marine mammals, and seabirds. This species is of great nutritional and economic value as a cheap but rich source of protein and fat. It is either consumed directly by humans or fed as fishmeal for aquaculture and farm animals. Despite its importance in the food basket, little is known about the molecular mechanisms involved in protein and lipid synthesis in this species. We collected nine tissues of *Sardina pilchardus* and reconstructed the transcriptome. In all, 198,597 transcripts were obtained, from which 68,031 are protein-coding. Quality assessment of the transcriptome was performed by back-mapping reads to the transcriptome and by searching for Single Copy Orthologs. Additionally, Gene Ontology and KEGG annotations were retrieved for most of the protein-coding genes. Finally, each library was quantified in terms of Transcripts per Million to disclose their expression patterns.

Keywords

Sardina pilchardus, European sardine, Transcriptome assembly, Annotation, Expression, Tissue quantification, Pathway, Gene Ontology

Specifications Table

Subject	Omics: Transcriptomics
Specific subject area	Transcriptomics, Genomics, Fisheries, Aquaculture
Type of data	Tables, Figures, FASTA Assembly, FASTQ read files
How data were acquired	Illumina HiSeq 2000 sequencing platform
Data format	Raw reads(FASTQ) Assembly (FASTA) Annotation (TSV) Quantification (TSV)
Parameters for data collection	Three sardines were collected by IFREMER during a scientific bottom trawl survey.
Description of data collection	Total RNA was collected from nine tissues: brain, eye, heart, kidney, liver, muscle, ovary, skin, and testes. Sequencing was performed using an Illumina HiSeq 2000, yielding single-stranded paired-end reads with a length of 101 bp. Reads were cleaned with Trimmomatic. Assembly was performed with Trinity. Assembly quality was assessed with Bowtie2 and BUSCO. Annotation was done with TransDecoder and Trinotate. Quantification was performed with kallisto and sleuth.
Data source location	IFREMER survey EVHOE 2015, 31-10-2015, Bay of Biscay, 47°18' N, 2°46' W
Data accessibility	Raw RNA-seq reads of *Sardina pilchardus* are deposited at ENA Bioproject PRJEB18441 https://www.ebi.ac.uk/ena/browser/view/PRJEB18441.The following tissues are available: brain (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925802), eye (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925803), heart (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925804), kidney (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925805), liver (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925806), muscle (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925807), ovary 1 (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925808), ovary 2 (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925809), skin (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925810),

	and testes (ERR5925802; https://www.ebi.ac.uk/ena/browser/view/ERR5925811) Supplementary data is available at Figshare under DOI 10.6084/m9.figshare.14617149 (https://doi.org/10.6084/m9.figshare.14617149.v1) .

Value of the Data

- We present the Illumina sequencing effort and *de novo* transcriptome assembly of *Sardina pilchardus*, an important small pelagic fish due to its nutritional, economic, and ecological value.
- This data will facilitate genome annotation and the discovery of genes of interest for the aquaculture industry. This resource could serve as the basis of a SNP chip that could differentiate the stocks of sardines across the Atlantic Ocean and the Mediterranean Sea.
- The transcriptome, annotation, and expression patterns can be used to study the genes and pathways involved in ω-3 fatty acid synthesis and storage.
- The tissue quantification can be used to perform an RT-qPCR of a transcript of interest, using the tissue in which we know the target gene is active.
- Comparative evolutionary studies can be done to unravel the phylogenetic relationship of the sardine within the Clupeiformes or other teleost species.
- Selection signatures can be identified by investigating functional differences between orthologous genes in sardines and other Clupeiformes species inhabiting different environments.

1. Data Description

This dataset contains the RNA-Seq analysis of nine tissues of *Sardina pilchardus*. Nine tissues from two female and one male sardines were dissected onboard and immersed immediately in RNAlater. Sequencing was performed using the Illumina HiSeq 2000 platform, yielding 56 million single-stranded paired-end reads of length 101 base pairs, a median quality value per sequence of 37, 5.6 million reads per sample on average, resulting in a total of 5.70 Gbp (Table 1). Reads were preprocessed with Trimmomatic, which slightly reduced the dataset to 98,09% of the reads, and the mean read length to 100.67 base pairs. Clean reads were assembled with Trinity. To measure the quality of the assembly, cleaned reads were back-mapped to the reference, and transcripts were searched for *Actinopterygii* Single-Copy Orthologs (SCOs). Transcripts were annotated with TransDecoder and Trinotate. Results of the sequencing effort and read cleaning are available in Table 1, while the ones of assembly, quality control and annotation are in Table 2. Figure 1 shows the most frequent Gene Ontology annotations received, and the coverage of the metabolome based on the KEGG annotations. Finally, each library was quantified with kallisto and prepared for differential downstream analysis with sleuth to obtain the expression patterns for each transcript in every tissue. The raw reads for the nine tissues of *Sardina pilchardus* have been deposited at the European Nucleotide Archive, under the umbrella project PRJEB18441, while each experimental run is deposited under accession numbers ERR5925802 to

ERR5925811 (Table 1). To our knowledge, this is one of the widest datasets not only in Clupeiformes but also in fish in general, only surpassed by the ones in (1). Supplementary data with the raw transcriptome assembly, predicted protein-coding sequences, transcript annotation and tissue quantification are available at Figshare under DOI 10.6084/m9.figshare.14617149. It includes: the assembled transcriptome (sd01-assembly.fasta), the predicted coding-sequences (sd02-transdecoder.cds), annotation (sd03-trinotate.tsv) and expression profiles per tissue (sd04-tpms.tsv).

Table 1
Summary of the read cleaning and backmapping of every library against the assembled reference.

Library	Sample	Accession number	Raw reads (M)	Trimmed reads (M)	Trimmed %	Trimmed Gbp	Mapped %
Brain	F1	ERR5925802	6,11	6,00	98,29	0,60	95,88
Eye	F1	ERR5925803	5,34	5,23	97,99	0,53	98,38
Heart	F1	ERR5925804	4,98	4,89	98,24	0,49	98,99
Kidney	M	ERR5925805	6,68	6,56	98,18	0,66	97,20
Liver	F1	ERR5925806	4,67	4,59	98,23	0,46	98,86
Muscle	F1	ERR5925807	5,31	5,24	98,67	0,53	98,21
Ovary 1	F1	ERR5925808	6,64	6,50	98,00	0,66	98,03
Ovary 2	F2	ERR5925809	6,57	6,41	97,60	0,65	98,05
Skin	M	ERR5925810	5,17	5,06	97,90	0,51	97,46
Testes	M	ERR5925811	5,04	4,93	97,84	0,50	97,09
Total			56,52	55,43	98,09	5,58	97,77

Sample: sample used, M for male, F1 and F2 for the females.
Raw: Original number of reads from the sequencer, in millions.
Clean: number of reads free of adapters and sequencing errors, in millions.
Clean %: Fraction of the original reads free of adapters and sequencing errors.
Clean Gbp: Total number of error-free bases, in giga base pairs.
Mapped %: Fraction of the trimmed reads that are back-mapped to the transcriptome.

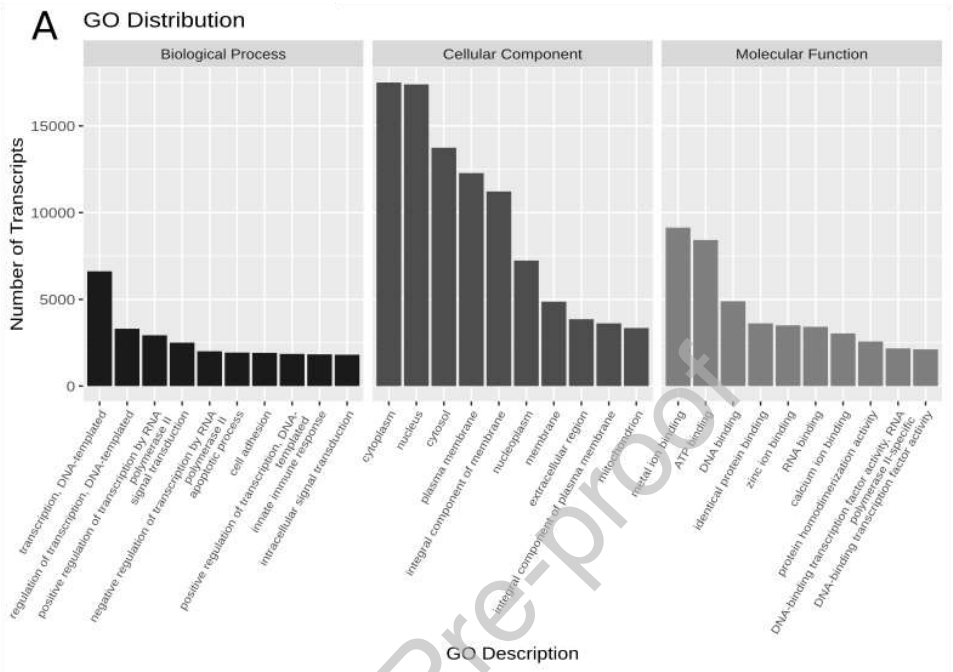
Table 2
Summary statistics of *de novo* transcriptome assembly, quality assessment, and annotation for *Sardina pilchardus* using nine tissues.

Assembly description	
Assembled transcripts	198597
Unigenes	149981
Assembly length (Mbp)	149.36
N10	3475
N30	2080
N50	1280
Average contig length	752.08
Longest contig length	10795
GC%	48,1
Quality Control	
Mapped reads	97,80%
Actinopterygii BUSCOs	4584
Complete, single copy	45,60%
Complete, duplicated	27,80%
Fragmented	11,60%
Missing	15,00%
Annotation	
Predicted ORF	68031
Complete proteins	24187
Contigs with match to SwissProt	67772
Contigs with GO term	66396
Contigs with PFAM domain	45154
Contigs with KEGG annotation	59254

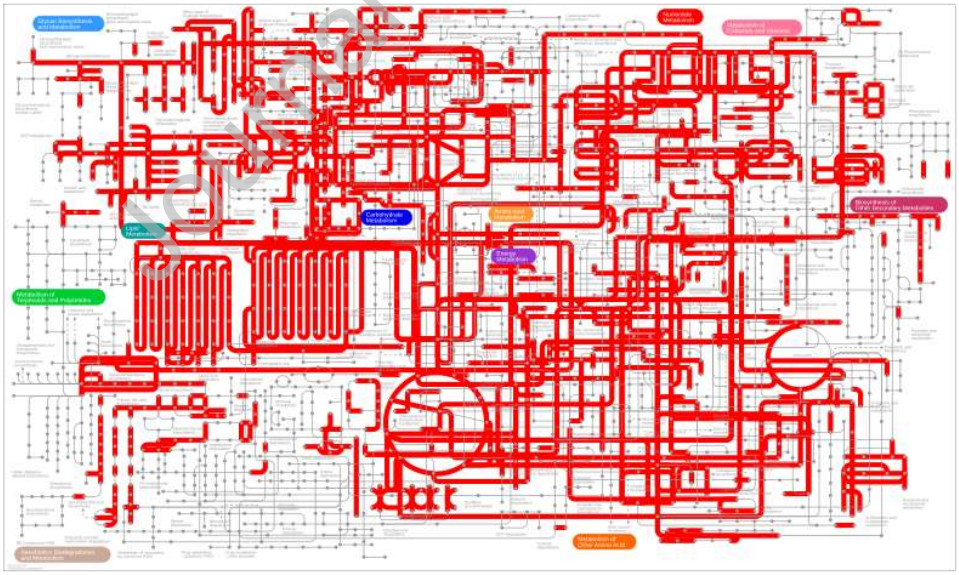
Figure 1

A. Gene Ontology annotation of the *Sardina pilchardus* transcriptome. The figure shows the top ten level 2 categories within the three principal categories

B. Expressed metabolome of *Sardina pilchardus* based on the KEGG annotation.



B



2. Experimental Design, Materials and Methods

2.1 Sampling strategy

Three individuals from the European Atlantic Ocean were collected by the IFREMER institute during the EVHOE scientific surveys (October 10th, 2015(2)). From these individuals, nine tissues (brain, eye, heart, kidney, liver, muscle, ovaries, skin, and testes) were dissected onboard, immediately immersed in RNAlater (Invitrogen), and stored at -20ºC until further processing.

2.2 RNA extraction, library construction, and sequencing

Total RNA from nine tissues (Table 1) and three individuals were extracted using TriZol® Reagent (Life Technologies) and quantified with Agilent 2100 Bioanalyzer combined with Agilent RNA 6000 Nano chips (Agilent Technologies, Inc.) at the Gene Expression Unit (SGIker) of the University of the Basque Country UPV/EHU. Samples with RNA integrity numbers (RIN) below 8 were immediately discarded. For every tissue, the sample with the highest RIN was used for sequencing. The exception was testes since there was only one male specimen, and ovary, where both samples were used. A multiplex sequencing library was prepared by labeling each sample with specific 10-mer barcoding oligonucleotides. The barcoded RNA-Seq libraries were sequenced using the Illumina HiSeq 2000 platform using one single lane. Sequencing reactions were performed with paired-end 101-bp and strand-specific protocol at the sequencing facility of the CNAG (Centre Nacional d'Anàlisi Genòmica, Barcelona, Spain). Base-calling was performed using the Illumina native software.

2.3 Read processing, assembly and quality control

Raw reads were processed with Trimmomatic v0.33 (3) using a gentle procedure to remove adapters and low-quality bases, using the parameters 'SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25'. The trimmed reads were assembled with Trinity (4), using default parameters with the exception that input reads were single-stranded to optimize the assembly. To understand the reliability of this assembly, a two-fold approach was used to study its completeness and representativeness. First, the transcriptome was analyzed by running BUSCO (5) against the *Actinopterygii* (ray-finned fishes) database. This software compares the transcriptome against a precomputed set of proteins conserved as Single-Copy Orthologs (SCOs) and returns how many of them are found, duplicated, fragmented or missing. Second, the representativeness of reference was obtained with Bowtie2 (6).

2.4 Functional annotation and quantification

Functional annotation of the transcriptome was performed with the execution of the protein prediction software TransDecoder v5.0.2 (4) followed by the annotation of both transcripts and proteins with Trinotate v3.0.2 (7).

TransDecoder translated each transcript into the six possible amino acid sequences and filtered out Open Reading Frames shorter than 300 nucleotides. Afterward, each candidate protein was queried against the SwissProt (8) and Pfam-A (9) databases (downloaded on 2018-10-22) and retained those hits with an E-value or domain noise cutoff less than or equal to 1e-5.

Subsequently, Trinotate was executed with default settings and using the same SwissProt and Pfam databases as before, and the same databases and threshold parameters for BLASTX, BLASTP, and hmmscan. Briefly, transcripts, predicted coding-sequences, and proteins are compared against the SwissProt and Pfam databases, and for each positive match, the source sequence inherits the annotation of its entry in its respective database. This way, sequences obtain Gene Ontology (10) and KEGG (11). Annotations were obtained for 55,781 proteins from at least one database. Figure 1 shows the Gene Ontology distribution of terms, and the parts of the metabolome covered, according to the KEGG annotation, and generated with the ggplot2 R package (12), and IPath3.0 (13), respectively.

Trimmed reads were pseudo-aligned and quantified with kallisto v0.44.0 (14) and normalized Transcript per Million counts were obtained with Sleuth v0.29.0 (15).

Ethics Statement

Research complies with the ARRIVE guidelines and was conducted in accordance with the EU directive 2010/63/EU. IFREMER research vessels are under the supervision of the French Ministry of Education and Research. A steering committee evaluates and approves the campaign program.

Funding Information

We gratefully acknowledge funding from the Basque Government through a predoctoral grant (PRE_2017_2_0169) and from the Basque University System research group IT1233-19, "Applied Genomics and Bioinformatics". We also acknowledge funding from the IFREMER institute and by FFP (France Filière Pêche) through the project CAPTAIN.

CRediT author statement

Jorge Langa: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation Resources, Data Curation, Writing - Original Draft, Visualization; **Iratxe Montes:** Conceptualization, Investigation, Resources; **Martin Huret:** Conceptualization, Resources, Writing - Review & Editing; **Darrell Conklin:** Conceptualization, Writing - Review & Editing, Supervision; **Andone Estonba:** Conceptualization, Methodology, Supervision, Writing - Review & Editing, Project administration, Funding acquisition.

Acknowledgments

The authors are thankful for the technical and human support provided by Fernando Rendo, Irati Miguel, and Irantzu Bernales from the Genomics Service (SGIker) at the UPV/EHU. We also thank the crew of the R/V THALASSA, as well as E. Duhamel and P. Gatti for the onboard sampling.

Declaration of Competing Interest

The authors declare that they have no competing financial interests, which could influence the work reported in this article.

References

1. Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, et al. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC Genomics. 2016 May 18;17(1):368.
2. Leaute J-P, Pawloski L, Salaun M. EVHOE 2015 cruise,Thalassa R/V. 2015 [cited 2021 May 25]; Available from: https://doi.org/10.17600/15002200
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
4. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Agosto;8(8):1494–512.
5. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.
6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.
7. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 2017 Jan 17;18(3):762–76.
8. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.
9. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–32.
10. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330–8.
11. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.
12. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2009. Available from: http://ggplot2.org
13. Darzi Y, Letunic I, Bork P, Yamada T. iPath3.0: interactive pathways explorer v3. Nucleic Acids Res. 2018 Jul 2;46(W1):W510–3.
14. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 May;34(5):525–7.
15. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017 Jul;14(7):687–90.

Part IV

Supplement: Articles under review

Appendix D

Article 4. Clupeiformes

Langa, J., Rueda, Y., Albaina, A., Huret, M., Conklin, D., & Estonba, A. (n.d.). Recurrent positive selection of lipid trafficking genes in Clupeiformes, Manuscript under review on Marine Biotechnology.

Article coauthored with Dr. Yuri Rueda. It is under review at Marine Biotechnology.

TABLE D.1: Quality Metrics for Marine Biotechnology in 2020.

Marine Biotechnology 2020 - Web of Science	
Category	Marine and Freshwater Biology
Impact Factor	3.619
Rank	14/110
Quantile	Q1

Marine Biotechnology 2020 - Scopus	
Category	Aquatic Science
CiteScore	5.2
Rank	20 / 224
Quantile	Q1

Recurrent positive selection of lipid trafficking genes in Clupeiformes

Running title: Positive selection in Clupeiformes

Authors

Jorge Langa[1*], Yuri Rueda[2*], Aitor Albaina[3], Martin Huret[4], Darrell Conklin[5,6], Andone Estonba[1]

Affiliations

1. Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Spain
2. Department of Physiology, Faculty of Medicine, University of the Basque Country UPV/EHU, 48940 Leioa, Spain
3. Department of Zoology and Animal Cell Biology, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Spain
4. IFREMER, STH/LBH, B.P. 70, Plouzané, 29280 France
5. Department of Computer Science and Artificial Intelligence, Faculty of Computer Science, University of the Basque Country UPV/EHU, San Sebastián, Spain
6. IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Keywords

Phylogenetics, Clupeiformes, Lipids, Positive selection, HDL, LDL

Corresponding author: Jorge Langa, jorgeeliseo.langa@ehu.eus

Abstract

Clupeiformes is one of the most important orders of fishes due to its ecological and economic importance. The plasticity that their genomes show explains the ubiquity of these species across the globe: they occupy tropical and temperate latitudes, they are typically marine but adaptable to freshwater. However, little is known about their concrete genetic makeover and evolutive strategies, or how these species have become one of the richest sources of omega-3 long-chain polyunsaturated fatty acids (ω-3 LC-PUFAs). Here we report the discovery of genes and families under positive selection in the Clupeiformes order. In general, we found positively selected portions of the genome related to mitochondria, ribosomes, lysosomes, caveolae, extracellular proteins, and CD molecules. Furthermore, we

observe positively selected genes associated mainly with apolipoproteins and *caveolae*, among others, indicating that these fishes have adapted their molecular machinery to efficiently store and transport fats across tissues. The herein applied methodology and the obtained results pave the way for further research into the evolutionary history of Clupeiformes, while also streamlining the study of another set of species, from raw RNA-Seq reads to tabular results.

Translated abstract

Clupeiformes es uno de los órdenes más importantes de peces debido a su importancia ecológica y económica. La plasticidad que sus genomas muestran explica la ubicuidad de estas especies alrededor del mundo: ocupan latitudes tropicales y templadas, típicamente son marinos, pero se adaptan al agua dulce. Sin embargo, poco se sabe de su composición genética y estrategias evolutivas, o cómo estas especies se han convertido en una de las fuentes más ricas de ácidos grasos poliinsaturados de cadena larga (ω-3 LC-PUFAs). Aquí reportamos el descubrimiento de genes y familias bajo selección positiva en el orden Clupeiformes. En general, descubrimos seleccionados positivamente porciones del genoma relacionados con mitocondrias, ribosomas, lisosomas, caveolas, proteínas extracelulares, y moléculas CD. Es más, observamos genes seleccionados positivamente asociados principalmente con apolipoproteínas y caveolas, entre otras, indicando que estos peces han adaptado su maquinaria molecular para almacenar y transportar eficientemente grasas entre los tejidos. La metodología aplicada aquí y los resultados obtenidos allanan el camino para estudios posteriores sobre la historia evolutiva de los Clupeiformes, a la vez que agiliza el estudio de cualquier otro conjunto de especies, desde lecturas brutas de RNA-Seq hasta resultados en forma de tabla.

1 Introduction

Clupeiformes is an order of ray-finned fishes (Teleostei) that contains approximately 400 species including anchovies, herrings, allis, shads, and sardines, among others (Bloom and

Egan, 2018). These species occupy a wide range of habitats across the globe, from tropical to temperate latitudes, and from marine to freshwater. Additionally, they are keystone species mediating between the plankton on the bottom and the predators on top of the trophic web. From an anthropocentric perspective, these fishes are important for their economic value due to their high content of lipids and proteins. According to the Food and Agriculture Organization (FAO), herrings, anchovies, and sardines supposed 24% across all reported fish catches in weight in 2018 (http://www.fao.org/fishery/statistics/global-capture-production/query/en). Due to overfishing and global warming, the stocks of these species are decreasing, and periodic collapses are reported with a subsequent fishing ban (ICES, 2018).

One of the reasons for the ubiquity of these species across the globe is the plasticity that their genomes show. These fishes inhabit tropical and temperate latitudes across the globe with some species showing a bipolar distribution such as in the genus *Engraulis* (e.g. (Grant et al., 2005)). Although they are typically marine, some live in brackish and even freshwater environments including anadromous species. Interestingly, even fully marine species such as European anchovy (*Engraulis encrasicolus*, Linnaeus, 1758), that can inhabit coastal waters from 0º to 60º at both hemispheres, are capable of successfully spawning at extremely contrasting salinity sites, from oceanic waters to river water plumes (e.g. (Motos et al., 1996)). Clupeiformes have also adapted to forage a wide variety of plankton from phytoplankton to mesozooplankton (Egan et al., 2018). The high diversity of this order across contrasting biotic and abiotic factors makes Clupeiformes a good candidate when looking for significant drivers of speciation, in terms of signatures of positive selection, in fishes.

Even though their huge economic and dietary importance, with dedicated fisheries worldwide due to their abundance and high nutritional value primarily linked to its high omega-3 long-chain polyunsaturated fatty acid (ω-3 LC-PUFA) content (e.g. (Machado et al., 2018)) and their beneficial effects (Burdge, 1998; Kim and Mendis, 2006; Lemaitre et al., 2003; Ruxton et al., 2004; Sidhu, 2003; Uauy and Valenzuela, 2000; Yokoyama et al., 2007), few genomic

resources exist for this group. To our knowledge, only two genome assemblies are available at Ensembl (*Clupea harengus*, Linnaeus, 1758, and *Denticeps clupeoides*, Clausen, 1959; (í Kongsstovu et al., 2019; Kersey et al., 2018; Rhie et al., 2021)) along with two drafts for *Sardina pilchardus* (Walbaum, 1792) (Machado et al., 2018; Louro et al., 2019). Interestingly, teleost fishes went under a whole-genome duplication (WGD) which occurred 320-350 million years ago. This initially duplicated the gene repertoire and, with time, gave new functions to some copies -neofunctionalization-, specialize some of their functions -subfunctionalization-, and, most of the time, deleted one of the copies (Nei and Roychoudhury, 1973; Takahata and Maruyama, 1979; Watterson, 1983; Force et al., 1999). This phenomenon not only makes it hard to assemble the genome and transcriptome but also to delimitate if different gene products are isoforms (i.e., are the effect of different RNA splicing patterns) or they come from paralogous genes. In this regard, until whole-genome assemblies do not become available for these fishes, the cost-effective approach to characterize protein-coding genes is to study the expressed messenger RNAs.

To face the environmental factors that perturb homeostasis, such as the contrasting temperature and salinity environments these species inhabit, the genome has either to ramp up mRNA transcription to counter the external changes by producing more proteins or to tailor mutations to synthesize more efficient proteins. To study the first strategy, the go-to method is to perform an RNA-Seq approach on two or more conditions to characterize the key genes that are under- and over-expressed. To address the second approach, phylogenetic tests have been developed to detect evolutionary differences across different species. Among these, the site test (Yang et al., 2000) was developed to detect positive selection in a limited number of sites of a protein. Posteriorly, the branch test (Yang, 1998) detected positive selection in a larger set of sites of a protein but over a limited amount of time. Finally, the branch-site test (Zhang et al., 2005) was presented to address the limitations of the two previous methods (Roux et al., 2014). Introduced in the CodeML module from PAML (Phylogenetic Analysis by Maximum Likelihood; (Yang, 2007), the three different tests have been used genome- and transcriptome-wide to identify genes under the

effect of positive selection (Roux et al., 2014; Ciezarek et al., 2016). The application of this procedure to every single gene and branch allows us to discover patterns of evolution without any hypothesis in mind, and therefore unbiased in a search of a goal in particular.

In this regard, we sailed out to discover genes under positive selection in Clupeiformes that could shed light on the biodiversity and evolutionary relationships of this group. To do so, we downloaded transcriptomic data from publicly available studies for twelve species, assembled them *de novo* when necessary, recovered the protein-coding sequences, clustered them into orthogroups, built the species phylogenetic tree, and tested each orthogroup for positive selection using the branch-site test. Then, we tested each Gene Ontology term, Reactome pathway, and HUGO Gene Nomenclature Consortium (HGNC) gene family for overrepresentation to discover also which modules are under strong positive selection. Additionally, we recorded which orthogroups are under selection in multiple branches at once.

2 Materials and Methods

2.1 Initial transcriptomes

To guide the transcript clustering step, reference transcriptomes were downloaded from Ensembl (Kersey et al., 2018) for teleosts *Astyanax mexicanus*, *Danio rerio*, *Gasterosteus aculeatus*, *Gadus morhua*, *Latimeria chalumnae*, *Lepisosteus oculatus*, *Oryzias latipes*, *Oreochromis niloticus*, *Poecilia formosa*, and *Xiphophorus maculatus*; and mammals *Homo sapiens* and *Mus musculus*.

Predicted transcriptomes for *D. clupeoides* and *C. harengus* were downloaded from their respective NCBI Genome entries (accession numbers available at Supplementary Table ST01). Assembled transcriptomes from *Konosirus punctatu*s (Temminck & Schlegel, 1846) and *Alosa pseudoharengus* (Wilson, 1811) (Pasquier et al., 2016) were downloaded from

ENA's Transcriptome Shotgun Assembly (TSA) website. For the remainder 8 species (*E. encrasicolus*, *Coilia nasus* (Temminck and Schlegel, 1846), *Clupea pallasii* (Valenciennes, 1847)*, Tenualosa ilisha* (Hamilton, 1822), *Alosa alosa* (Linnaeus, 1758)*, Brevoortia tyrannus* (Latrobe, 1802)*, Sardinops sagax* (Jenyns, 1842)*, S. pilchardus* (Walbaum, 1792)), transcriptomes were built by *de novo* assembly of previously published RNA-Seq datasets ((Roberts et al., 2012; Eldem et al., 2015; Iv et al., 2017; Zhu et al., 2017; Richards et al., 2018; Divya et al., 2019; Langa et al., 2021); Table 1 shows the species description, source of samples, and sequencing instruments used; ENA accession numbers available at Supplementary Table ST01).

For the 454 reads dataset in *C. pallasii*, basecalls were corrected and converted to FASTQ with PyroBayes 0.9 (Quinlan et al., 2008), trimmed, and cleaned for contamination with SnoWhite 2.0.3 (Dlugosch et al., 2013), and assembled with gsAssembler 2.9 (Roche Ltd.). In the case of Illumina datasets, reads were processed with Trimmomatic 0.36 (Bolger et al., 2014) by removing characters at the 5' and 3' ends with quality below 20 (LEADING:20 TRAILING:20), removing reads with average quality below 30 (AVGQUAL:30) and with length less than 32 nucleotides (MINLEN:32). Reads were normalized *in silico* with khmer 2.0 (Crusoe et al., 2015), except for *S. pilchardus* and *S. sagax*, due to the low number of bases sequenced. For each experiment, four hash tables 4 GB in size each were constructed to normalize the reads to a 20x coverage and eliminate erroneous and highly covered 32-mers. Finally, transcriptomes were assembled with Trinity 2.2.0 (Haas et al., 2013) allowing a maximum of 40 GB when counting *k*-mers. Both Snakemake (Köster and Rahmann, 2012) Illumina and 454 assembly pipelines are available online at GitHub and Figshare (see Data Accessibility section).

As a quality control measure, the completeness of each transcriptome was assessed by searching Single Copy Orthologs with BUSCO 3.0.2 (Waterhouse et al., 2018). Briefly, this program analyzes an input set of sequences and performs BLAST (Camacho et al., 2009) and HMMER (Eddy, 2011) searches against a chosen dataset (*Actinopterygii*, ray-finned

fishes) of known Single-Copy Orthologs (SCOs) derived from the OrthoDB 8 database (Kriventseva et al., 2019). In the end, it reports which SCOs were found as single-copy, duplicated, fragmented, or missing.

2.2 Coding Sequence and protein prediction

Transcripts were inspected with TransDecoder 5.0.2 (Haas et al., 2013) to predict the coding sequences (CDS) and protein they encode while also removing 5' and 3'-UTR regions. This procedure was applied to all transcriptomes (Ensembl, NCBI Genome, ENA TSA, and *de novo* assembled). As the procedure requires using a single genetic code, mitochondrial transcripts were discarded in favor of mitogenomic data already published (see section 2.5). TransDecoder, in a first step, searches for putative ORFs with a minimum length of 300 nucleotides (100 amino-acids), and then performs homology searches with Diamond 0.9.29 (Buchfink et al., 2015) and hmmscan (Eddy, 2011) against the SwissProt ("UniProt," 2019) and Pfam-A (El-Gebali et al., 2019) databases, respectively. Finally, CDS and protein sequences are predicted. To automate the procedure, another Snakemake pipeline was built, available at GitHub and Dryad (see Data Accessibility section).

2.3 Orthology Inference and alignment refinement

To cluster and align the coding sequences obtained, we based our approach on the ones used in (Roux et al., 2014) and (Ciezarek et al., 2016), which in turn are based on the procedures from Ensembl Compara (Herrero et al., 2016) and Selectome (Proux et al., 2009; Moretti et al., 2014).

Protein coding sequences were clustered with CD-HIT-EST 4.8.1 (Li and Godzik, 2006) with a similarity threshold of 99.5% to remove within-species redundancy. Redundancy-free CDS for all 24 species were clustered with OrthoFinder 2.3.3 (Emms and Kelly, 2019) to predict orthogroups (families of homologous sequences - putative genes) across all species. This program first performs an all-vs-all Diamond search of the translated CDS sequences, then

normalized the results through the bit scores of the searches, taking into account sequence lengths and phylogenetic distances, and then performed clustering of the graph induced by the Diamond bit scores with MCL 14.137 (van Dongen and Abreu-Goodger, 2012).

To separate paralogous genes inside each orthogroup with certainty, we executed the methods of (Yang and Smith, 2014). This approach consists of two rounds of protein realignment with MAFFT 7.464 (Katoh and Standley, 2013), column trimming with pxclsq (from the phyx package 1.01; (Brown et al., 2017)), tree inference with RAXML-NG 0.9.0 (Kozlov et al., 2019) using the WAG protein evolution model (Whelan and Goldman, 2001), trimming of the tree' tips that had an absolute length of 2 or a relative length of 10 times its sister tip (trim_tips.py), removal of tips from the same species with fewer characters while also removing the paraphyletic ones (mask_tips_by_taxonID_transcripts.py), and removal of deep paralogs (tips with a branch length greater than 0.5; cut_long_internal_branches.py). After these two rounds of refinement, high-quality orthologs were predicted with the Root-to-Tip method (prune_paralogs_RT.py), which takes into account gene duplication events, such as the WGD that occurred in teleosts. Mammals and non-clupeid fish species were used as outgroups and removed from the CDS/protein clusters and phylogenetic trees.

Since alignment errors are an important source of false positives in the search for Positive Selection (Löytynoja, 2014; Redelings, 2014), a more stringent procedure consisting of two more rounds of refinement was applied. First, alignments were performed with M-Coffee 11.0.8 (Wallace et al., 2006), which aligned independently each orthogroup with Muscle 3.8.31 (Edgar, 2004), MAFFT 7.464 (Katoh and Standley, 2013), T-Coffee 11.0.8 (Notredame et al., 2000), and kalign 2.04 (Lassmann et al., 2009). Then, alignments were evaluated, and columns with a score below 9 out of 9 were removed. Finally, proteins were back-translated to CDS, columns (codons) with occupancy below 50% were removed with pxclsq, and rows (transcripts) rich in gaps were removed too with MaxAlign 1.1 (Gouveia-Oliveira et al., 2007) using default settings. These trimmed sequences were reprocessed then a second time.

2.4 Phylogenetic Bayesian Tree Construction

Due to the redundancy of the genetic code, multiple codons encode the same amino acid. Four-fold degenerate sites are the positions in the third base of codon alignments that produce the same amino acid no matter what mutation occurs in that position. Therefore, a phylogeny built on these positions should be free from positive selection constraints (Eyre-Walker and Keightley, 1999; Nachman and Crowell, 2000). Four-fold degenerate sites were extracted from all the orthogroups previously aligned and were concatenated to form a supermatrix, requiring each orthogroup to contain at least four taxa. Then, columns were removed if they had occupancy of less than 50% with pxclsq, and converted to PHYLIP format. ModelTest-NG (Darriba et al., 2020) was executed to see which evolution model fitted best our sequence alignment. Then, the filtered supermatrix and the fittest model were fed to RAxML-NG to obtain the maximum likelihood tree, performing also 1,000 bootstrap replicates. Finally, ExaBayes 1.5 (Aberer et al., 2014) was run to obtain the Bayesian phylogenetic species tree, using the previous tree as the starting tree, four independent MCMC independent runs, with 3 coupled chains and 1,000,000 generations each, and sampling every 500 generations. Finally, sdsf and postProcParams were run to ensure that the split frequencies, scale reduction factors, and effective sample sizes were close to zero, one, and 200, respectively. From the bootstrap replicates, a consensus unrooted tree was generated, and from this one, a rooted version was generated with pxrr using *D. clupeoides* as the outgroup.

2.5 Detection of Positive Selection

The branch-site test (Zhang et al., 2005) from CodeML, part of the PAML's 4.9j package (Yang, 2007) was the one used as the core of this manuscript. Instead of running CodeML directly, we used the Python package ETE3 3.1.1 (Huerta-Cepas et al., 2016) due to its convenience: it acts as an interface to CodeML through the command line, avoiding designing one control file per branch and orthogroup, parsing the very verbose outputs of CodeML, and being able to plot each alignment, signaling p-

values and ω ratios for the different models used. To carry over the analysis, each orthogroup had to have a minimum of two species in both the foreground and background branches.

To do so, the aforementioned method studies the ratio between nonsynonymous (dN) to synonymous (dS) substitutions, denoted by ω (ω = dN/dS). On the one hand, the null model assumes that the evolution speed ω_f of the foreground branch is strictly less than one ($\omega_f < 1$). On the other hand, the alternative hypothesis allows four situations: $0 < \omega_{background}, \omega_{foreground} < 1$ (codons are conserved), $\omega_{foreground} = \omega_{background} = 1$ (codons are evolving neutrally), $\omega_{foreground} > 1 = \omega_{background}$ (the foreground is under positive selection while the background is evolving neutrally), and $\omega_{foreground} > 1 > \omega_{background}$ (the foreground is selected, while the background is fixed). For each branch and ortholog analyzed, a likelihood ratio test (LRT) is used to compare whether the alternate model fits better than the null hypothesis by comparing it to a χ^2 test. Due to the sensitivity of the method to initial conditions (Yang and dos Reis, 2011), for every ortholog and branch, ETE3 was executed three times with different starting values for ω_0 (0.5, 1.0, 1.5). Orthologs and branches were considered putatively under selection if the tests were found significant (p-value < 0.05) in the three different starting points.

Since alignment error is a key source of false-positives (Markova-Raina and Petrov, 2011; Redelings, 2014), putatively selected orthogroups were processed again with an even stricter approach: Guidance2 2.02 (Sela et al., 2015), a codon-aware probabilistic aligner that has shown to produce low rates of false positives compared to previous approaches. This program was executed using PRANK v.170427 (Löytynoja, 2014), and running 100 bootstraps. Low-quality positions in the alignments were removed with TrimAl 1.2 (Capella-Gutiérrez et al., 2009), using the automated feature, and taxa rich in gaps were removed too with MaxAlign. Finally, a second step of detection of positive selection was performed, this time with FastCodeML 1.3.0 (Valle et al., 2014), a faster implementation of the branch-site

tests of CodeML, using also the different starting ω_0. As in ETE3, both foreground and background branches had the requirement of a minimum of two species.

Mitochondrial genes were analyzed and processed similarly. The 13 mitochondrial genes for all clupeoid species but one (*S. sagax*) were downloaded from NCBI Gene and separated into 13 different FASTA files. They were directly aligned with GUIDANCE2, trimmed with TrimAl, filtered with MaxAlign, and searched for positive selection with ETE3, using this time the vertebrate mitochondrial genetic code.

All p-values from the two methods (ETE3 and FastCodeML) and three starting points were merged across all orthologs and all branches analyzed and corrected with the Benjamini-Hochberg method. All ortholog-branches with all six corrected p-values under 0.05 were considered to be under Positive Selection. To avoid working with multiple adjusted p-values per starting point, for each orthogroup and branch analyzed, the maximum of the adjusted p-values were reported.

Since synonymous-site saturation (SSS) may influence the reliability of the branch-site test (Gharib and Robinson-Rechavi, 2013; Roux et al., 2014), CodeML's free-ratio (b_free) and one-ratio models (M0) were run again with ETE3 in the selected orthogroups-branches. Orthogroup-branches with dS > 1 in both models were considered to be under SSS, and therefore possible false positives.

2.6 Orthogroup annotation and enrichment

Every transcript in every high-quality orthogroup was searched for homology against the Ensembl's Zebrafish transcriptome with DIAMOND's BLASTX implementation, matching each clupeid transcript to a single zebrafish transcript, and therefore a zebrafish gene via lowest e-value. Then, a gene symbol was assigned to each orthogroup via a majority rule of the symbols associated with its constituent transcripts. Given that Clupeiformes genes may or may not have been conserved after the teleost WGD event, a suffix of the form "-n|m" was

added to the associated gene symbol, indicating that such an orthogroup is the copy number n out of m. . Orthogroups that did not match any zebrafish gene were simply given the symbol “unknown” followed by a suffix. Tables with the equivalences between the Zebrafish genes and their Gene Ontology (Ashburner et al., 2000; “The Gene Ontology Resource,” 2019) and Reactome (Jassal et al., 2020) annotations were downloaded from Ensembl’s Biomart (Kinsella et al., 2011). Then, each orthogroup inherited the annotations of its corresponding Zebrafish gene. Additionally, gene family annotations and human to zebrafish gene orthologs were downloaded from the HGNC database (Braschi et al., 2019) and Ensembl’s Biomart, and each clupeoid orthogroup inherited the gene family annotation of their zebrafish ortholog, which in turn inherited it from its human equivalent. When possible, HGNC families were subdivided into subfamilies for a more granular analysis. For example, the human “Aldo-keto reductase family” could be subdivided into subfamilies “Aldo-reductase family 1” to “Aldo-keto reductase family 7”. This division of the gene families can discover subfamilies under positive selection while its superfamily is not.

Enrichment of Gene Ontology, Reactome, and HGNC clusters was analyzed with the Bioconductor (Huber et al., 2015) ClusterProfiler 3.16.0 package (Yu et al., 2012). This R (R Core Team, 2020) package performs enrichment of any custom set of terms with the “enrich” function and provides multiple forms to visualize the results. For the three collections of terms, we used as the foreground those orthogroups that were selected and had an annotation, and as the background list the set of orthogroups that were annotated. In both cases, terms associated with between 5 and 500 terms (minGsSize = 5, maxGsSize = 500) were kept to avoid very specific and non-specific categories, respectively. For each term, a Fisher exact test was performed, and all p-values were corrected by the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). Additional data processing and visualization were done with the R Tidyverse metapackage (Wickham et al., 2019).

Finally, we tracked the number of times an orthogroup was positively selected, to obtain which ones are under strong evolutive pressure, even divergence.

3 Results

3.1 Sequence assembly, protein-coding prediction

For Illumina experiments, between 5 and 66 Gbp were assembled into 107,804 to 885,281 transcripts (Table 2). *T. ilisha* obtained the lowest number of assembled transcripts (107,804) because of the low sequencing depth (6 Gbp) and being composed of a single library (liver). In the same sense, as expected the *C. pallasii* 454 transcriptome yielded much fewer transcripts (19,004) than Illumina transcriptomes.

3.2 Protein-coding prediction and initial filtering

Transdecoder predicted which sequences had a coding potential. In total, between 5,788 (*C. pallasii*) and 206,368 (*E. encrasicolus*) transcripts were predicted to be protein-coding, discarding between 6% (*C. harengus*) to 84% (*S. sagax*) of the input transcripts. We note that the transcriptomes from reference species were the ones that obtained the highest rate of conversion to CDS (94% in *C. harengus* and 87% for *D. clupeoides*), due to being the product of an annotation of a genome assembly rather than a *de novo* transcriptome assembly. Outside those cases, the conversion rates fell to a range between 14% (*S. sagax*) to 40% (*K. punctatus* and *A. pseudoharengus*).

Results from BUSCO returned an approximation of the completeness of the transcriptome, revealing that between 5.4% to 97.6% of the Actinopterygii Single-Copy Orthologs (SCOs) were present in the studied transcriptomes (Table 2 and Figure 2). *C. pallasii* and *S. sagax* obtained the lowest results (5.4% and 11.7%) due to the relatively low sequencing effort. Nonetheless, those transcriptomes were kept to analyze the Clupeinae and *Sardina+Sardinops* branches for positive selection. Outside the species with the genome assembly, *S. pilchardus* obtained the highest counts for SCOs, followed by *A. pseudoharengus.* We hypothesize the high results for *S. pilchardus (7*3%) are due to the

width of tissues sampled, although *A. alosa* was sequenced with a similar tissue width but with ten times the depth, and obtained only 32% of SCOs and, similar results were obtained for *A. pseudoharengus* where the only reported tissue were gills.

Finally, CD-HIT-EST eliminated the redundancy of the coding sequences, reducing the transcriptomes between 0.16% (*K. punctatus*) to 33.03% (*D. clupeoides*; Table 2). Due to the lack of a study associated with *K. punctatus*, we suspect that this published transcriptome was already clustered. Transcriptomes inferred from assembled genomes were the ones that benefited more from the clustering step (*C. harengus* and *D. clupeoides*; 26.13% and 33.03% respectively), due to the specificity of Ensembl's gene prediction algorithms.

3.3 Orthogroup clustering, refinements, and Species Tree construction

OrthoFinder initially discovered 26,020 orthogroups composed of 936,769 transcripts, which after the different refinement steps were reduced to 19,914 high-quality orthogroups and 97,383 transcripts. From these refined orthogroups, a supermatrix of 4-fold degenerate sites was built, composed of 15,877 loci and 1,293,781 characters and an overall occupancy of 41.98%, with a widely different contribution from the species: 3.87% from *C. pallasii* to 72.11% of *S. pilchardus*. A maximum-likelihood tree was computed with RAxML-NG, with the evolution model proposed by ModelTest-NG (“TVM+G4”). This ML tree was used as the starting tree for ExaBayes. The resulting bayesian tree is the one depicted in Figure 1 (rooted using *D. clupeoides* as outgroup). The posterior probability of this tree is 1.0 and all 100 bootstrap trees supported this topology. This tree separates Clupeinae from the subfamily composed of Alosinae and Dorosomatinae.

3.4 Testing for selection

After building the Clupeiformes species tree, we tested for positive selection the eight branches marked in red in Figure 1 out of the possible 22. We required alignments to contain at least 2 species both in the background and the foreground, discarding immediately the analysis of the 12 terminal branches, the Cluepidei one, and the one that separates Brevoortia from Alosa. In the end, we tested for selection the Engraulidae and Clupeidae families, Clupeinae, Dorosomatinae and Alosinae subfamilies, and the aggregation of the last two (Alosinae+Dorosomatinae), and genera *Clupea*, and the combinations of *Sardina*+*Sardinops*, and *Alosa*+*Brevoortia.* Given the completeness results from BUSCO, branches Clupeinae and Sardina+Sardinops had difficulties meeting that criteria but were analyzed nonetheless. Also, given that transcriptomes from *C. harengus* and *D. clupeoides* are derived from genome assemblies, almost all branches have assured at least one species in the background.

In total, 15,822 orthogroups were susceptible to be tested for selection, ranging from 483 in *Clupeinae*, to 11,349 in "Alosinae+Dorosomatinae" due to the high number of species present (Table 3). After the two rounds of realignment, testing for selection, and p-value correction, 918 orthogroups (5.8%) were found to be under positive selection in any of the 8 branches analyzed. On a per-branch basis, from 23 (Clupeinae, 4.76%) to 331 orthogroups (Alosinae and Alosinae+Dorosomatinae, 3.01% and 2.92%, respectively) were found to be positively selected. Analysis of the speed of synonymous changes determined that 4 orthogroups (out of the 918) were under SSS, minimizing the presence of false positives.

3.5 Annotation

The 19,914 high-quality orthogroups were annotated by aligning them with DIAMOND against all the zebrafish cDNA set from Ensembl. From these 19,914 orthogroups, 18,960 matched 14,116 *D. rerio* genes, while the remainder 954 did not show sufficient homology, suggesting that these 954 orthogroups correspond to genes present in the Clupeiformes and

D. rerio last common ancestor, but lost in the evolution of the latter. A non-exhaustive search of these orthogroups suggests that 295 of these unknown orthogroups match at least one human gene, while the remainder still lack an annotation. Also, due to the many-to-one assignment of orthogroups to *D. rerio* genes, we inferred the duplication of some genes. Among them, we found *FO704673.1* (12 copies, annotated as an endonuclease), *zgc:100868* (10 copies, a peptidase hydrolase), *bptf* (10 copies, subcomponent of the nuclease remodeling factor), *BX546500.1* (10 copies, another endonuclease), and *CR848040.3* (10 copies, a reverse transcriptase domain-containing protein; Supplementary Table ST03). Given the clustering steps done with CD-HIT, OrthoFinder, and the subsequent stringent filtering procedure by (Y. Yang & Smith, 2014), we minimized false positive duplications due to artifacts of alternative splicing and kept these orthogroups for the rest of the analysis.

Regarding the unknown orthogroups, 44 out of 965 were found under positive selection (4.45%). Additionally, much of the genes also found under selection have been identified (for example, *si:dkeyp-84f3.5*, *BX901897.1*, or *CABZ01068246.1*), but little is known about them: they are successfully cloned mRNAs, or genes predicted *ab initio* whose name relates to the contig they belong to.

3.6 Overselected GO terms, Reactome pathways, and HGNC gene families

When taking into account the evolution across all eight branches, we obtained eight GO terms to be enriched after p-value correction (Table 5 and Figure 3). Six of these terms are related to three organelles in particular: two to the mitochondrion and the electron transport chain (ETC; “BP: mitochondrial electron transport, cytochrome c to oxygen” and “BP: ATP synthesis coupled proton transport”), three to the ribosome (“BP: translation”, “CC: ribosome”, and “MF: structural constituent of ribosome”), and one to the lysosome (“CC: lysosome”). The two remaining terms were associated with caveolae (“CC: caveola”), and a

non-specific set of genes whose products are secreted outside the cell (“CC: extracellular region”).

Concerning the Reactome pathways, two clusters with four pathways each were obtained, both referring, as with GO terms, to the translation process ("Translation", "Mitochondrial translation", "Mitochondrial translation elongation" and "Mitochondrial translation termination") and energy production processes, focusing again on the ETC ("The citric acid (TCA) cycle and respiratory electron transport", "Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins", "Respiratory electron transport", and "Complex I biogenesis", Table 6 and Figure 4).

Finally, the HGNC gene family and subfamily enrichment revealed overrepresentation of both complexes I and V of the ETC (“NADH:ubiquinone oxidoreductase supernumerary subunits”, “Mitochondrial complex V: ATP synthase subunits”), and also “CD molecules”, “Apolipoproteins”, and the “Solute Carrier Family 25”, also known as mitochondrial carriers (Table 7 and Figure 5).

Therefore, taking together the GO, Reactome, and HGNC results, Clupeiformes have been shaped through positive selection on six elements: mitochondria, ribosomes, lysosomes, caveolae, extracellular proteins, and CD molecules.

Additionally, genes selected multiple times across the different branches were studied. Among them, we found orthogroups that either have been found in Zebrafish but are of unknown function (*CR846080.1-2|2*, *si:ch211-15j1.5*, or *si:ch211-207i1.2-3|7*)), or that seem to be missing completely in that genome (*unknown-565|965*, *unknown-422|965*, or *unknown-185|965*; Table 8 and Supplementary Table ST04). Apart from this, we obtained multiple apolipoproteins: *apoa1a-1|2* (depicted in Figure 6), *apoa1a-2|2*, *apoa2*, *apoeb-1|2*, and *apooa*. Another example of selection in multiple branches is the *mylipa-1|3* gene, which regulates cholesterol uptake (Lindholm et al., 2009; Zelcer et al., 2009).

4. Discussion

4.1 Pipeline construction

This paper has presented the results of a transcriptome-wide scan for positively selected genes in twelve Clupeiformes species. To do so, we designed a Snakemake pipeline, available at GitHub and deposited at Figshare, based on previously described methods (Proux et al., 2009; Moretti et al., 2014; Roux et al., 2014; Ciezarek et al., 2016; Herrero et al., 2016). This pipeline, which takes as input the assembled transcripts, performs the necessary steps to cluster them into the high-quality orthogroups, computes the Bayesian phylogenetic species tree derived from four-fold degenerate sites, and searches which orthogroups are under positive selection through CodeML's branch-site test, presented in (Zhang et al., 2005), and implemented in PAML (Z. Yang, 2007). This pipeline automates the procedure, leaving to the user the task of providing the assembled transcriptomes and the branches to analyze, and provides the results as tables. Additionally, due to Snakemake's ability to manage conda environments, the installation of all required tools to perform the analysis is simplified and automated too, and even containerization of the entire procedure, ensuring not only reproducibility but wide applicability. Finally, special attention has been given to parallelization. Whenever possible all alignment, tree inference, and evolutionary tests are run in parallel, and the analysis of each branch can be done in different nodes, leveraging the power of HPC clusters. Therefore, we provide an easy to install and ready-to-use tool to search for signals of positive selection from any number of RNA-Seq experiments. Complementarily, we provide two independent pipelines to perform *de novo* transcriptome assemblies with Illumina and 454 technologies, and another one to annotate them with TransDecoder and Trinotate (see Data Accessibility section).

4.2 Gene clustering, phylogenetic relationships, and search for positive Selection

The results presented in this document highlight interesting aspects of molecular evolution within the Clupeiformes order. We clustered the initial million and a half transcripts into a total of high-quality 19,914 protein-coding orthogroups, which matched a significant fraction of the 25,592 protein-coding genes in Zebrafish (GRCz10). Given the triple filtering procedure used (CD-HIT to remove redundancy within each species, resolution of gene trees in OrthoFinder, and posterior rounds of refinements), the probability of falsely duplicated genes is very low. Also, we were not able to retrieve a Zebrafish annotation from 965 orthogroups, suggesting that they likely are teleost genes lost in the evolution of Zebrafish (Supplementary Table ST02). A preliminary Diamond BLASTP search resulted in 282 out of the 952 unknown orthogroups with homology to the human proteome (data not shown).

All branches from the Bayesian phylogenetic tree in Figure 1 obtained a 100% bootstrap support and agree with the dated trees from previous works (Lavoué et al., 2014; Egan et al., 2018; Bloom & Egan, 2018), derived from a wide number of species but limited to a handful of nuclear and mitochondrial sequences. In turn, it disagrees with the ones presented in (Nelson et al., 2016) and (Whitehead et al., 1985), derived from morphological characters alone, in which the *Tenualosa* genus is placed within Alosinae instead of Dososomatinae. While the access to RNA-Seq data is becoming more common due to the advances in sequencing technologies, the elucidation of phylogenetic relationships with the herein presented approach, or a similar one, will likely become a new standard to complete traditional methods (Ciezarek et al., 2016; S. Li et al., 2018; Roux et al., 2014; Spalink et al., 2018; Wang et al., 2017). Although with a limited number of species, present data confirming the most recent DNA-based phylogeny reconstructions suggest its huge potential.

4.3 Categories of genes under selection

Our analysis of overrepresented GO terms, Reactome pathways, and HGNC families found that evolution has shaped the genomes of Clupeiformes in, at least, six putative ways.

(1) The ETC within mitochondria, where complexes I (NADH:ubiquinone oxidoreductase), IV (cytochrome c oxidase), and V (ATP Synthase) were significantly overrepresented in terms of GO, pathways, and HGNC families. No protein in complexes II (succinate dehydrogenase) and III (cytochrome bc_1) was positively selected in any branch. Additionally, mitochondrial carriers were selected too according to the analysis of HGNC gene families.

(2) The ribosomes and the translation processes were selected too, both nuclear and mitochondrial, with significantly more selected genes in the latter.

(3) The lysosome was overrepresented according to GO, although no clear pattern in the genes selected was found. Notably, we find selected the *npc1* gene, responsible for cholesterol excretion, together with *npc2*, although the latter lies outside this organelle.

(4) The *caveolae* GO term was selected, signaling cavins *cavin1b-1*|*3*, *cavin1b-2*|*3*, *cavin2b*, and *zgc:172270* and caveolins *cav1* and *cav3*. Although lying outside the "Caveolae" Gene Ontology term, we found positively selected *pacsin2*, essential to caveola assembly.

(5) The cluster of differentiation (CD) molecules appeared overrepresented in HGNC. They are cell-surface proteins used to characterize leukocytes and other immune system-relevant cells by immunophenotyping (Chan et al., 1988). They act as cell adhesion proteins, ligands, and receptors, and therefore participate in cell signaling (Table 7 and Figure 5).

(6) The wide term of extracellular protein-coding genes was also overrepresented. Although it is very generic, a pattern can be observed. Some genes were found related to immune and

inflammatory responses (b2ml-1|2, ccl20a.3, ccl34b.3, cxcl11.1, cxcl32b.1-2|2, il10, il1b), and signal transduction (ecm1a-2|2, ecm1b, gdf3, igf2b, il10, thpo). The most outstanding set of genes in this group are apolipoproteins that are discussed in more detail along with other genes involved in lipid metabolism.

4.4 Lipid trafficking genes

Across our results, we observe multiple times, in terms of Gene Ontologies, HGNC families, and gene descriptions, the presence of genes associated with lipid and lipoprotein biology. The GO term *caveola* (Table 5) and HGNC family *apolipoproteins* (Table 7) are the gene groups significantly selected, although several other genes related to lipids in general, and cholesterol and fatty acids, in particular, can be found among selected genes (Supplementary Table 3).

Apolipoproteins are the protein components of lipoproteins, which are aggregates of lipids and proteins needed to transport water-insoluble fats between different tissues. Although mammalian apolipoproteins have been mostly studied, the similarity in apolipoprotein nature and distribution between fish and mammals has been long described (Babin & Vernier, 1989). According to our results, selected apolipoproteins were ApoA-I and ApoA-II (*apoa1a-1|2*, *apoa1a-2|2*, *apoa2*; with the first and the third selected in three branches, and the latter with some of the lowest p-values; Table 4 and Table 8), ApoD (*apoda.2*), ApoE (*apoeb-1|2*), and ApoO (*apooa*).

Although present also in other lipoprotein types, ApoA-I and ApoA-II are the first and second most abundant protein components of high-density lipoproteins (HDL), while ApoD is a multi-ligand lipid carrier able to form heterodimers with ApoA apolipoproteins (Rassart et al., 2020). One of the main functions of HDL is to perform reverse cholesterol transport (RCT), i.e., the transport of excess cholesterol from peripheral tissues to the liver. In the liver, HDLs enter caveolae and interact there with the scavenger receptor BI (SCARB1, also known as SR-BI) which extracts esterified cholesterol from them, (Pilch et al., 2011; Liu et al., 2013;

Zanoni et al., 2018). Gene families *caveola* and *apolipoproteins* being selected suggests that RCT might be a process of evolutionary interest in the Clupeiformes. The protein that links together both, SR-BI, was present in all but one (*C. pallasii*) of the transcriptomes, however, the requirement of a high-quality alignment to avoid false positives resulted in the inability to even test this gene. Therefore, we were not able to test for positive selection of the gene encoding SR-BI (*scarb*).

In addition to RCT, our results suggest that processes related to cholesterol supply to peripheral cells have been of evolutionary interest for Clupeiformes. One of the main cholesterol sources for peripheral tissues are low density lipoproteins (LDL), which transport hepatic cholesterol to peripheral tissues. Cells in need of cholesterol express LDL receptor (LDLR) on their surface and internalize LDL particles via clathrin-mediated endocytosis. LDL-containing endosomes fuse with lysosomes (where esterified cholesterol is hydrolyzed) and cholesterol is transported to the cytosol in a process mediated by proteins Niemann-Pick disease type C1 (NPC1) and NPC2 (Zanoni et al., 2018). Clathrin light chain B (*cltb*), *npc1* and *npc2* are among the selected genes. E3 ubiquitin ligase myosin regulatory light chain-interacting protein (*mylipa-1*|*3*), which induces the degradation of LDL receptors (Lindholm et al., 2009; Zelcer et al., 2009), and paraoxonase 2 (*pon2-2*|*2*), which prevents LDL and HDL oxidation (Aviram et al., 1998; Mackness et al., 1993), are shown to be selected too (Supplementary Table ST03). The other main cholesterol source is *de novo* synthesis and the enzyme catalyzing the last step of cholesterol synthesis (*dhcr7*) happens to be positively selected too.

Overall, trafficking of cholesterol-rich lipoproteins and cholesterol supply seem to be of evolutionary interest for this group of animals. Cholesterol is a precursor to other steroid molecules (e.g. hormones, vitamin D and bile acids) and determines physical properties of the plasma membrane including toughness, permeability and fluidity. The possible relationship between those cholesterol-related factors and the evolutionary success leading

to the ubiquity of Clupeiformes remains unveiled and might be of great interest for future research.

Sardines, anchovies and herrings are known for being a rich source of long chain polyunsaturated fatty acids (LC-PUFAs), in particular the ω-3 type. The consumption of ω-3 LC-PUFAs is linked to a lesser probability of cardiovascular disease in humans (Lemaitre et al., 2003; Yokoyama et al., 2007; Jain et al., 2015), lesser tissue inflammation (Uauy & Valenzuela, 2000; Ruxton et al., 2004), and development of nervous (Burdge, 1998), reproductive (Sidhu, 2003) and visual systems (Kim & Mendis, 2006). Briefly, ω-3 and ω-6 LC-PUFAs are synthesized by the repetitive elongation and desaturation of saturated fatty acids by ELOVL and FADS proteins, respectively. On the one hand, due to the loss of the FADS1 gene in teleosts, Clupeiformes and many other fishes cannot synthesize arachidonic and eicosapentaenoic fatty acids (Machado et al., 2018; Garrido et al., 2019), the building blocks of eicosanoid hormones. Therefore, their presence is exclusively due to their planktonic dietary intake.

Anyhow, our results do show a selection of a set of genes related to fatty acid metabolism. Those include genes involved in fatty acid elongation (*elovl4b*, *hsd17b12a*, *hsd17b12b*) and desaturation (*fads2*). Fatty acid uptake from blood (*apoeb-1|2*, *slc27a4-1|2*) and their conversion to the metabolically available fatty acyl-coenzyme A (*acot18-2|3*, *acsf2-1|2*, *acsl5*, *acsm3*) are also reflected in the selected gene list. Due to detergent properties of fatty acids, cells usually store them esterified in lipids such as phospholipids and triglycerides or they bind to carrier proteins; genes related to the synthesis (*cds1*, *cers2b*, *fa2h-2|2*, *gpd1c*) or degradation (*lipia*) of such lipids and fatty acid carriers (*acbd4*, *fabp11a-1|2*) are also present in our gene list. Finally, we have found that several genes related to mitochondrial or peroxisomal degradation of fatty acids are also evolutionarily selected (*acadm-2|2*, *ech1*, *eci1*, *ehhadh*, *hadhb*, *slc25a20*) (Supplementary Table 3).

What is the exact role of these processes in the accumulation of ω-3 LC-PUFAs coming from dietary intake? What is their evolutionary impact on Clupeiformes? A hint could be the cyclical food availability, especially as latitude increases. European anchovies and sardines are more energy-dense in higher latitudes in order to survive the winter, when plankton is scarce or unavailable (Gatti et al., 2018; Huret et al., 2019). These species require rapid storage when food is available to overwinter and mobilization during that period, to then start spending energy on reproduction over the spawning season. Therefore it would be interesting to study the interaction between the adaptation within species across latitudes, the seasonal environment and the reproduction cycle, which is more challenging polewards.

In conclusion, we fully automated the procedure to identify genes under positive selection based on massive RNA-seq datasets, and applied it to the 19,914 orthogorups of twelve Clupeiformes species. Briefly, we clustered the transcripts into orthogroups and built a Bayesian phylogenetic tree. Based on this tree, each one of the orthogroups was tested for positive selection in the principal branches within Clupeiformes. Finally, the set of positively selected orthogroups were analyzed for overrepresentation of GO terms, Reactome pathways, and HGNC gene families. This study shows that evolution has acted on the mechanisms of lipid trafficking through the selection of caveolins, cavins, apolipoproteins and other genes related to lipid trafficking and metabolism. It remains to study if evolution has adapted this strategy only in the fishes herein explored, if it is common to all Clupeiformes, or if it has also happened in other teleosts rich in ω-3 type fatty acids. We expect this document to shed light on the evolutionary relationships within Clupeiformes and on their remarkable ecological success, and to open the door to test new hypotheses into their evolutionary biology and ecology. Finally, the herein developed bioinformatics could be directly applied to any other organism where RNAseq data is available to infer phylogenies and/or look for genes and families under selection.

Acknowledgments

The authors thank the University of the Basque Country for funding the publishing of this article as Open Access. We would like to acknowledge the technical and human support provided by IZO-SGI SGIker (UPV/EHU, MICINN, GV/EJ, ESF). JL was funded by the Department of Education from the Basque Government grant PRE_2017_2_0169, and by the Applied Genomics and Bioinformatics research group of the Basque Country (grants IT558-10 and IT1233-19). JL would also like to thank the Computational Phylogenetics group from the University of Lausanne, especially Dr. Martha Serrano and Professor Nicolas Salamin. Additionally, we would like to thank Dr. Ibón Cancio, from the Plentzia Marine Station.

Data Accessibility

The pipelines and datasets here used and analyzed are available at GitHub and Figshare:

- Illumina assembly pipeline: https://github.com/jlanga/smsk_khmer_trinity and https://doi.org/10.6084/m9.figshare.15117708
- 454 assembly pipeline: https://github.com/jlanga/smsk_454 and https://doi.org/10.6084/m9.figshare.15117657
- Transcriptome annotation https://github.com/jlanga/smsk_trinotate and https://doi.org/10.6084/m9.figshare.15118347
- Clustering, species tree construction and positive selection pipeline: https://github.com/jlanga/smsk_selection and https://doi.org/10.6084/m9.figshare.15120243
- Input transcriptomes, CDS and protein predictions: https://doi.org/10.6084/m9.figshare.15147492
- Final analysis to reproduce the tables and figures: https://doi.org/10.6084/m9.figshare.15121110
- Accession numbers for raw data and the transcriptome assemblies are available in Supplementary Table 1.

Author Contributions

J.L. and A.A. conceived and designed the study. J.L. wrote all the pipelines and analyzed the data. J.L., Y.R. and A.A. interpreted the results. J.L., A.A. and Y.R. wrote the manuscript with assistance of M.H., D.C. and A.E.

References

Aberer, A. J., Kobert, K., & Stamatakis, A. (2014). ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Molecular Biology and Evolution*, *31*(10), 2553–2556. https://doi.org/10.1093/molbev/msu236

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*, 25–29. https://doi.org/10.1038/75556

Aviram, M., Rosenblat, M., Bisgaier, C. L., Newton, R. S., Primo-Parmo, S. L., & La Du, B. N. (1998). Paraoxonase inhibits high-density lipoprotein oxidation and preserves its functions. A possible peroxidative role for paraoxonase. *Journal of Clinical Investigation*, *101*(8), 1581–1590. https://doi.org/10.1172/JCI1649

Babin, P. J., & Vernier, J. M. (1989). Plasma lipoproteins in fish. *Journal of Lipid Research*, *30*(4), 467–489.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bloom, D. D., & Egan, J. P. (2018). Systematics of Clupeiformes and testing for ecological limits on species richness in a trans-marine/freshwater clade. *Neotropical Ichthyology*, *16*(3). https://doi.org/10.1590/1982-0224-20180095

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

Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., Yates, B., & Bruford, E. (2019). Genenames.org: The HGNC and VGNC resources in 2019. *Nucleic Acids Research*, *47*(D1), D786–D792. https://doi.org/10.1093/nar/gky930

Brown, J. W., Walker, J. F., & Smith, S. A. (2017). Phyx: Phylogenetic tools for unix. *Bioinformatics*, *33*(12), 1886–1888. https://doi.org/10.1093/bioinformatics/btx063

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

Burdge, G. C. (1998). The role of docosahexaenoic acid in brain development and fetal alcohol syndrome. *Biochemical Society Transactions*, *26*(2), 246–251. https://doi.org/10.1042/bst0260246

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Chan, J. K. C., Ng, C. S., & Hui, P. K. (1988). A simple guide to the terminology and application of leucocyte monoclonal antibodies. *Histopathology*, *12*(5), 461–480. https://doi.org/10.1111/j.1365-2559.1988.tb01967.x

Ciezarek, A. G., Dunning, L. T., Jones, C. S., Noble, L. R., Humble, E., Stefanni, S. S., & Savolainen, V. (2016). Substitutions in the Glycogenin-1 Gene Are Associated with the Evolution of Endothermy in Sharks and Tunas. *Genome Biology and Evolution*, *8*(9), 3011–3021. https://doi.org/10.1093/gbe/evw211

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., Fenton, J., Fenzl, T., Fish, J., Garcia-Gutierrez, L., Garland, P., Gluck, J., González, I., Guermond, S., Guo, J., … Brown, C. T. (2015). The khmer software package: Enabling efficient nucleotide sequence analysis. *F1000Research*. https://doi.org/10.12688/f1000research.6924.1

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294. https://doi.org/10.1093/molbev/msz189

Divya, B. K., Mohindra, V., Singh, R. K., Yadav, P., Masih, P., & Jena, J. K. (2019). Muscle transcriptome resource for growth, lipid metabolism and immune system in Hilsa shad, Tenualosa ilisha. *Genes & Genomics*, *41*(1), 1–15. https://doi.org/10.1007/s13258-018-0732-y

Dlugosch, K. M., Lai, Z., Bonin, A., Hierro, J., & Rieseberg, L. H. (2013). Allele Identification for Transcriptome-Based Population Genomics in the Invasive Plant Centaurea solstitialis. *G3: Genes, Genomes, Genetics*, *3*(2), 359–367. https://doi.org/10.1534/g3.112.003871

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Comput Biol*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Egan, J. P., Bloom, D. D., Kuo, C.-H., Hammer, M. P., Tongnunui, P., Iglésias, S. P., Sheaves, M., Grudpan, C., & Simons, A. M. (2018). Phylogenetic analysis of trophic niche evolution reveals a latitudinal herbivory gradient in Clupeoidei (herrings,

anchovies, and allies). *Molecular Phylogenetics and Evolution*, *124*, 151–161. https://doi.org/10.1016/j.ympev.2018.03.011

Eldem, V., Zararsız, G., Erkan, M., & Bakir, Y. (2015). De novo assembly and comprehensive characterization of the skeletal muscle transcriptomes of the European anchovy (Engraulis encrasicolus). *Marine Genomics*, *20*, 7–9. https://doi.org/10.1016/j.margen.2015.01.001

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Eyre-Walker, A., & Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. *Nature*, *397*(6717), 344–347. https://doi.org/10.1038/16915

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545.

Garrido, D., Kabeya, N., Betancor, M. B., Pérez, J. A., Acosta, N. G., Tocher, D. R., Rodríguez, C., & Monroig, Ó. (2019). Functional diversification of teleost Fads2 fatty acyl desaturases occurs independently of the trophic level. *Scientific Reports*, *9*(1), 11199. https://doi.org/10.1038/s41598-019-47709-0

Gatti, P., Cominassi, L., Duhamel, E., Grellier, P., Le Delliou, H., Le Mestre, S., Petitgas, P., Rabiller, M., Spitz, J., & Huret, M. (2018). Bioenergetic condition of anchovy and sardine in the Bay of Biscay and English Channel. *Progress in Oceanography*, *166*, 129–138. https://doi.org/10.1016/j.pocean.2017.12.006

Gharib, W. H., & Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Molecular Biology and Evolution*, *30*(7), 1675–1686. https://doi.org/10.1093/molbev/mst062

Gouveia-Oliveira, R., Sackett, P. W., & Pedersen, A. G. (2007). MaxAlign: Maximizing usable data in an alignment. *BMC Bioinformatics*, *8*(1), 312. https://doi.org/10.1186/1471-2105-8-312

Grant, W. S., Leslie, R. W., & Bowen, B. W. (2005). Molecular genetic assessment of bipolarity in the anchovy genus Engraulis. *Journal of Fish Biology*, *67*(5), 1242–1265. https://doi.org/10.1111/j.1095-8649.2005.00820.x

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

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (2016). Ensembl comparative genomics resources. *Database*, *2016*(bav096). https://doi.org/10.1093/database/bav096

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Huret, M., Tsiaras, K., Daewel, U., Skogen, M. D., Gatti, P., Petitgas, P., & Somarakis, S. (2019). Variation in life-history traits of European anchovy along a latitudinal gradient: A bioenergetics modelling approach. *Marine Ecology Progress Series*, *617*, 95–112.

í Kongsstovu, S., Mikalsen, S.-O., Homrum, E. í, Jacobsen, J. A., Flicek, P., & Dahl, H. A. (2019). Using long and linked reads to improve an Atlantic herring (Clupea harengus) genome assembly. *Scientific Reports*, *9*(1), 17716. https://doi.org/10.1038/s41598-019-54151-9

ICES. (2018). *Report of the Working Group on Southern Horse Mackerel, Anchovy and Sardine (WGHANSA)*. *ICES CM 2018/ACOM:17*, 605 pp. http://www.ices.dk/community/groups/Pages/WGhansa.aspx

Iv, F. J. Z., Rana, S. B., Alvi, Z. A., Zhang, Z., Murphy, W., & Bentivegna, C. S. (2017). De Novo Assembly and Analysis of the Testes Transcriptome from the Menhaden, Bervoortia tyrannus. *Fisheries and Aquaculture Journal*, *8*(1), 1–8. https://doi.org/10.4172/2150-3508.1000186

Jain, A. P., Aggarwal, K. K., & Zhang, P.-Y. (2015). Omega-3 fatty acids and cardiovascular disease. *European Review for Medical and Pharmacological Sciences*, *19*(3), 441–445.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., Kumar, N., Liu, Z., Maurel, T., Moore, B., McDowall, M. D., Maheswari, U., Naamati, G., Newman, V., Ong, C. K., … Yates, A. (2018). Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*, *46*(D1), D802–D808. https://doi.org/10.1093/nar/gkx1011

Kim, S.-K., & Mendis, E. (2006). Bioactive compounds from marine processing byproducts – A review. *Food Research International*, *39*(4), 383–393. https://doi.org/10.1016/j.foodres.2005.10.010

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., & Flicek, P. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation*, *2011*, bar030. https://doi.org/10.1093/database/bar030

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811. https://doi.org/10.1093/nar/gky1053

Langa, J., Huret, M., Montes, I., Conklin, D., & Estonba, A. (2021). Transcriptomic dataset for Sardina pilchardus: Assembly, annotation, and expression of nine tissues. *Manuscript Submitted for Publication.*

Lassmann, T., Frings, O., & Sonnhammer, E. L. L. (2009). Kalign2: High-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Research*, *37*(3), 858–865. https://doi.org/10.1093/nar/gkn1006

Lavoué, S., Konstantinidis, P., Chen, W.-J., Konstantinidis, P., & Chen, W.-J. (2014, March 14). - *Progress in Clupeiform Systematics*. Biology and Ecology of Sardines and Anchovies; CRC Press. https://doi.org/10.1201/b16682-6

Lemaitre, R. N., King, I. B., Mozaffarian, D., Kuller, L. H., Tracy, R. P., & Siscovick, D. S. (2003). n−3 Polyunsaturated fatty acids, fatal ischemic heart disease, and nonfatal myocardial infarction in older adults: The Cardiovascular Health Study. *The American*

Journal of Clinical Nutrition, *77*(2), 319–325. https://doi.org/10.1093/ajcn/77.2.319

Li, S., Zhong, M., Dong, X., Jiang, X., Xu, Y., Sun, Y., Cheng, F., Li, D., Tang, K., Wang, S., Dai, S., & Hu, J.-Y. (2018). Comparative transcriptomics identifies patterns of selection in roses. *BMC Plant Biology*, *18*(1), 1–12. https://doi.org/10.1186/s12870-018-1585-x

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lindholm, D., Bornhauser, B. C., & Korhonen, L. (2009). Mylip makes an Idol turn into regulation of LDL receptor. *Cellular and Molecular Life Sciences*, *66*(21), 3399–3402. https://doi.org/10.1007/s00018-009-0127-y

Liu, X., Suo, R., Xiong, S.-L., Zhang, Q.-H., & Yi, G.-H. (2013). HDL drug carriers for targeted therapy. *Clinica Chimica Acta*, *415*, 94–100. https://doi.org/10.1016/j.cca.2012.10.008

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

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology (Clifton, N.J.)*, *1079*, 155–170. https://doi.org/10.1007/978-1-62703-646-7_10

Machado, A. M., Tørresen, O. K., Kabeya, N., Couto, A., Petersen, B., Felício, M., Campos, P. F., Fonseca, E., Bandarra, N., Lopes-Marques, M., Ferraz, R., Ruivo, R., Fonseca, M. M., Jentoft, S., Monroig, Ó., Da Fonseca, R. R., & C. Castro, L. F. (2018). "Out of the Can": A Draft Genome Assembly, Liver Transcriptome, and Nutrigenomics of the European Sardine, Sardina pilchardus. *Genes*, *9*(10), 485. https://doi.org/10.3390/genes9100485

Mackness, M., Arrol, S., Abbott, C., & Durrington, P. (1993). Protection of low-density lipoprotein against oxidative modification by high-density lipoprotein associated paraoxonase. *Atherosclerosis*, *104*(1–2), 129–135. https://doi.org/10.1016/0021-9150(93)90183-U

Markova-Raina, P., & Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Research*, *21*(6), 863–874. https://doi.org/10.1101/gr.115949.110

Moretti, S., Laurenczy, B., Gharib, W. H., Castella, B., Kuzniar, A., Schabauer, H., Studer, R. A., Valle, M., Salamin, N., Stockinger, H., & Robinson-Rechavi, M. (2014). Selectome update: Quality control and computational improvements to a database of positive selection. *Nucleic Acids Research*, *42*(D1), D917–D921. https://doi.org/10.1093/nar/gkt1065

Motos, L., Uriarte, A., & Valencia, V. (1996). The spawning environment of the Bay of Biscay anchovy (Engraulis encrasicolus L.). *Scientia Marina*, *60*.

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304.

Nei, M., & Roychoudhury, A. K. (1973). Probability of Fixation and Mean Fixation Time of an Overdominant Mutation. *Genetics*, *74*(2), 371–380.

Nelson, J. S., Grande, T. C., & Wilson, M. V. H. (2016). *Fishes of the World: Nelson/Fishes of the World*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119174844

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

Pasquier, J., Cabau, C., Nguyen, T., Jouanno, E., Severac, D., Braasch, I., Journot, L., Pontarotti, P., Klopp, C., Postlethwait, J. H., Guiguen, Y., & Bobe, J. (2016). Gene evolution and gene expression after whole genome duplication in fish: The PhyloFish database. *BMC Genomics*, *17*(1), 368. https://doi.org/10.1186/s12864-016-2709-z

Pilch, P. F., Meshulam, T., Ding, S., & Liu, L. (2011). Caveolae and lipid trafficking in adipocytes. *Clinical Lipidology*, *6*(1), 49–58.

Proux, E., Studer, R. A., Moretti, S., & Robinson-Rechavi, M. (2009). Selectome: A database of positive selection. *Nucleic Acids Research*, *37*(suppl_1), D404–D407. https://doi.org/10.1093/nar/gkn768

Quinlan, A. R., Stewart, D. A., Strömberg, M. P., & Marth, G. T. (2008). Pyrobayes: An improved base caller for SNP discovery in pyrosequences. *Nature Methods*, *5*(2), 179–181. https://doi.org/10.1038/nmeth.1172

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. https://www.R-project.org/

Rassart, E., Desmarais, F., Najyb, O., Bergeron, K.-F., & Mounier, C. (2020). Apolipoprotein D. *Gene*, *756*, 144874. https://doi.org/10.1016/j.gene.2020.144874

Redelings, B. (2014). Erasing Errors due to Alignment Ambiguity When Estimating Positive Selection. *Molecular Biology and Evolution*, *31*(8), 1979–1993. https://doi.org/10.1093/molbev/msu174

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Richards, D. J., Renaud, L., Agarwal, N., Starr Hazard, E., Hyde, J., & Hardiman, G. (2018). De Novo Hepatic Transcriptome Assembly and Systems Level Analysis of Three Species of Dietary Fish, Sardinops sagax, Scomber japonicus, and Pleuronichthys verticalis. *Genes*, *9*(11), 521. https://doi.org/10.3390/genes9110521

Roberts, S. B., Hauser, L., Seeb, L. W., & Seeb, J. E. (2012). Development of Genomic Resources for Pacific Herring through Targeted Transcriptome Pyrosequencing. *PLOS ONE*, *7*(2), e30908. https://doi.org/10.1371/journal.pone.0030908

Roux, J., Privman, E., Moretti, S., Daub, J. T., Robinson-Rechavi, M., & Keller, L. (2014). Patterns of Positive Selection in Seven Ant Genomes. *Molecular Biology and Evolution*, *31*(7), 1661–1685. https://doi.org/10.1093/molbev/msu141

Ruxton, C. H. S., Reed, S. C., Simpson, M. J. A., & Millington, K. J. (2004). The health benefits of omega-3 polyunsaturated fatty acids: A review of the evidence. *Journal of Human Nutrition and Dietetics*, *17*(5), 449–459. https://doi.org/10.1111/j.1365-277X.2004.00552.x

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015). GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, *43*(W1), W7–W14. https://doi.org/10.1093/nar/gkv318

Sidhu, K. S. (2003). Health benefits and potential risks related to consumption of fish or fish oil. *Regulatory Toxicology and Pharmacology*, *38*(3), 336–344. https://doi.org/10.1016/j.yrtph.2003.07.002

Spalink, D., Stoffel, K., Walden, G. K., Hulse-Kemp, A. M., Hill, T. A., Van Deynze, A., & Bohs, L. (2018). Comparative transcriptomics and genomic patterns of discordance in Capsiceae (Solanaceae). *Molecular Phylogenetics and Evolution*, *126*, 293–302. https://doi.org/10.1016/j.ympev.2018.04.030

Takahata, N., & Maruyama, T. (1979). Polymorphism and loss of duplicate gene expression: A theoretical study with application of tetraploid fish. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(9), 4521–4525.

The Gene Ontology Resource: 20 years and still GOing strong. (2019). *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Uauy, R., & Valenzuela, A. (2000). Marine oils: The health benefits of n-3 fatty acids. *Nutrition*, *16*(7), 680–684. https://doi.org/10.1016/S0899-9007(00)00326-9

UniProt: A worldwide hub of protein knowledge. (2019). *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Valle, M., Schabauer, H., Pacher, C., Stockinger, H., Stamatakis, A., Robinson-Rechavi, M., & Salamin, N. (2014). Optimization strategies for fast detection of positive selection on phylogenetic trees. *Bioinformatics*, *30*(8), 1129–1137. https://doi.org/10.1093/bioinformatics/btt760

van Dongen, S., & Abreu-Goodger, C. (2012). Using MCL to Extract Clusters from Networks. In J. van Helden, A. Toussaint, & D. Thieffry (Eds.), *Bacterial Molecular Networks: Methods and Protocols* (pp. 281–295). Springer. https://doi.org/10.1007/978-1-61779-361-5_15

Wallace, I. M., O'Sullivan, O., Higgins, D. G., & Notredame, C. (2006). M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *34*(6),

1692–1699. https://doi.org/10.1093/nar/gkl091
Wang, K., Hong, W., Jiao, H., & Zhao, H. (2017). Transcriptome sequencing and phylogenetic analysis of four species of luminescent beetles. *Scientific Reports*, *7*(1), 1814. https://doi.org/10.1038/s41598-017-01835-9
Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319
Watterson, G. A. (1983). On the Time for Gene Silencing at Duplicate Loci. *Genetics*, *105*(3), 745–766.
Whelan, S., & Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851
Whitehead, P. J. P., Nelson, G. J., & Thosaporn Wongratana. (1985). *Clupeoid fishes of the world (suborder Clupeoidei): An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, shads, anchovies, and wolfherrings*. United Nations Development Programme : Food and Agriculture Organization of the United Nations.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686
Yang, Y., & Smith, S. A. (2014). Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution*, *31*(11), 3081–3092. https://doi.org/10.1093/molbev/msu245
Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, *15*(5), 568–573. https://doi.org/10.1093/oxfordjournals.molbev.a025957
Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088
Yang, Z., & dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, *28*(3), 1217–1228. https://doi.org/10.1093/molbev/msq303
Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A.-M. K. (2000). Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics*, *155*(1), 431–449.
Yokoyama, M., Origasa, H., Matsuzaki, M., Matsuzawa, Y., Saito, Y., Ishikawa, Y., Oikawa, S., Sasaki, J., Hishida, H., Itakura, H., Kita, T., Kitabatake, A., Nakaya, N., Sakata, T., Shimada, K., & Shirato, K. (2007). Effects of eicosapentaenoic acid on major coronary events in hypercholesterolaemic patients (JELIS): A randomised open-label, blinded endpoint analysis. *The Lancet*, *369*(9567), 1090–1098. https://doi.org/10.1016/S0140-6736(07)60527-3
Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118
Zanoni, P., Velagapudi, S., Yalcinkaya, M., Rohrer, L., & von Eckardstein, A. (2018). Endocytosis of lipoproteins. *Atherosclerosis*, *275*, 273–295. https://doi.org/10.1016/j.atherosclerosis.2018.06.881
Zelcer, N., Hong, C., Boyadjian, R., & Tontonoz, P. (2009). LXR Regulates Cholesterol Uptake Through Idol-Dependent Ubiquitination of the LDL Receptor. *Science*, *325*(5936), 100–104. https://doi.org/10.1126/science.1168974
Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution*, *22*(12), 2472–2479. https://doi.org/10.1093/molbev/msi237
Zhu, G., Wang, L., Tang, W., Wang, X., & Wang, C. (2017). Identification of olfactory receptor genes in the Japanese grenadier anchovy Coilia nasus. *Genes & Genomics*,

39(5), 521–532. https://doi.org/10.1007/s13258-017-0517-8

Tables

Table 1 - Species list

List with the species used for the analysis of this paper. Accession numbers for every sequencing run or assembly are available in Supplementary Table ST01

Suborder	Family	Subfamily	Species	Taxon id	Source	Tissues
Denticipitoidei	Denticipitidae	-	*Denticeps clupeoides*	299321	NCBI Annotation	-
Clupeoidei	Engraulidae	Engraulinae	*Engraulis encrasicolus*	184585	Illumina HiSeq 2000	Muscle
Clupeoidei	Engraulidae	Coilinae	*Coilia nasus*	365059	Illumina GAII / HiSeq 2000 / HiSeq 2500	Liver
Clupeoidei	Clupeidae	Clupeinae	*Clupea pallasii*	476916	Roche 454 GS FLX Titanium	Liver, testes
Clupeoidei	Clupeidae	Clupeinae	*Clupea harengus*	7950	NCBI Annotation	-
Clupeoidei	Clupeidae	Dorosomatinae	*Konosirus punctatus*	365056	ENA TSA	Muscle, liver, gill, heart, kidney, swim bladder, and sexual gland
Clupeoidei	Clupeidae	Dorosomatinae	*Tenualosa ilisha*	373995	Illumina HiSeq 2500	Liver
Clupeoidei	Clupeidae	Alosinae	*Brevoortia tyrannus*	224708	Illumina HiSeq 2000	Testes
Clupeoidei	Clupeidae	Alosinae	*Alosa alosa*	278164	Illumina HiSeq 2000	Brain, liver, gills, heart, muscle, liver, kidney, bones, intestine, ovary, and testes.
Clupeoidei	Clupeidae	Alosinae	*Alosa pseudoharengus*	34774	ENA TSA	Gill
Clupeoidei	Clupeidae	Alosinae	*Sardina pilchardus*	27697	Illumina HiSeq 2000	brain, ovary, liver, kidney, skin, testes, muscle, heart, eye
Clupeoidei	Clupeidae	Alosinae	*Sardinops sagax*	298279	Illumina GAIIx	Liver

Table 2 - Assembly and filtering

Assembly, TransDecoder, and CD-HIT results and presence of each species in the orthogroups tested for selection. The number of high-quality orthogroups is 19,914. BUSCO categories consist of complete (C), either Single copy (S) or duplicated (D), fragmented (F), and missing (M).

Species	Gbp	Transcripts	CDS	Busco Statistics	CD-HIT	Orthogroups
D. clupeoides	-	59645	51825	C:97.6%[S:48.4%,D:49.2%],F:1.1%,M:1.3%	35228	9125
E. encrasicolus	15	719059	206368	C:60.4%[S:37.8%,D:22.6%],F:18.8%,M:20.8%	180065	10983
C. nasus	55	885281	185428	C:35.1%[S:19.1%,D:16.0%],F:27.9%,M:37.0%	147342	8628
C. pallasii	0,6	19004	5788	C:5.4%[S:5.1%,D:0.3%],F:9.0%,M:85.6%	5588	643
C. harengus	-	46203	43379	C:93.0%[S:53.2%,D:39.8%],F:1.6%,M:5.4%	32042	13254
K. punctatus	-	69974	28028	C:54.6%[S:52.1%,D:2.5%],F:14.9%,M:30.5%	27982	8933
T. ilisha	6	107804	27355	C:29.5%[S:24.9%,D:4.6%],F:18.0%,M:52.5%	25395	5538
B. tyrannus	18	266785	58794	C:32.1%[S:23.8%,D:8.3%],F:24.9%,M:43.0%	51959	6904
A. alosa	66	734830	160367	C:31.8%[S:17.4%,D:14.4%],F:24.9%,M:43.3%	129407	7529
A. pseudoharengus	-	216529	86639	C:65.7%[S:43.2%,D:22.5%],F:15.8%,M:18.5%	81903	10127
S. pilchardus	6	198597	67993	C:73.2%[S:45.7%,D:27.5%],F:11.5%,M:15.3%	55803	13028
S. sagax	5	196984	27300	C:11.7%[S:8.9%,D:2.8%],F:21.6%,M:66.7%	24913	2691

Table 3 - Branches analyzed

Table with the names of the branches analyzed, species and families they contain, the number of orthogroups analyzed, how many of them were selected and how many of them were under synonymous site saturation (SSS).

Branch name	N. species	Valid alignments	Under selection	% selected	Selected and SSS	% SSS
Engraulidae	2	5130	192	3,74%	4	2,08%
Clupeinae	2	496	24	4,84%	0	0,00%
Dorosomatinae	2	2970	77	2,59%	0	0,00%
Sardina+Sardinops	2	1989	47	2,36%	0	0,00%
Alosa+Brevoortia	3	6787	148	2,18%	0	0,00%
Alosinae	5	11004	367	3,34%	1	0,27%
Alosinae+Dorosomatinae	7	11360	369	3,25%	2	0,54%
Clupeidae	9	8561	152	1,78%	0	0,00%
Total		15832	1376	8,69%	6	0,44%

Table 4 - Top-scoring genes

List with the top 20 orthogroups ranked by the maximum corrected p-value between the three tests carried out in fastcodeml. When the same orthogroup is selected in multiple branches, multiple p-values are shown. Seven mitochondrial genes appear on the top of this list. Note that Zebrafishes' gene symbols already contain letters a and b to differentiate between duplicated genes

Symbol	Description	Orthogroup ID	Max. adjusted p-value	Phylogroup	Ensembl ID
trip4	thyroid hormone receptor interactor 4	OG0008035_1_1.inclade1.ortho1	0.00E+00	*Alosa+Brevoortia*	ENSDARG00000098074.2
krt18a.1	keratin 18a, tandem duplicate 1	OG0001033_1_1.inclade1.ortho1	2.22e-15, 2.44e-15	Alosinae, *Alosa+Brevoortia*	ENSDARG00000018404.10
BX284638.2-1|3	wu:fb95e10	OG0011449_1_1.inclade1.ortho1	1.23E-12	Dorosomatinae	ENSDARG00000116871.1
si:ch211-140m22.7-2|2	si:ch211-140m22.7	OG0004030_2_1.inclade1.ortho1	1.56E-11	*Alosa+Brevoortia*	ENSDARG00000077967.6
cylda-4|4	cylindromatosis (turban tumor syndrome), a	OG0010158_1_1.inclade1.ortho1	2.21E-11	Alosinae	ENSDARG00000060058.9
PRRC2B-2|2	proline-rich coiled-coil 2B	OG0016888_1_1.unrooted-ortho	2.31E-10	Alosinae	ENSDARG00000079639.4
col10a1b	collagen, type X, alpha 1b	OG0002050_1_1.inclade2.ortho1	2.43E-09	Alosinae	ENSDARG00000101535.2
col4a2	collagen, type IV, alpha 2	OG0000218_1_1.inclade3.ortho1	5.08e-09, 6.80e-09	Alosinae+Dorosomatinae, Dorosomatinae	ENSDARG00000104110.2
si:ch211-209f23.6-2|3	si:ch211-209f23.6	OG0000081_2_1.inclade2.ortho1	5.40e-09, 5.40e-09, 5.40e-09	Alosinae+Dorosomatinae, *Alosa+Brevoortia*, Alosinae	ENSDARG00000077309.3
tmprss5-3|3	transmembrane serine protease 5	OG0015295_1_1.inclade1.ortho1	1.03E-08	*Alosa+Brevoortia*	ENSDARG00000087717.4
apoa2	apolipoprotein A-II	OG0012288_1_1.inclade1.ortho1	3.05e-08, 3.05e-08, 3.05e-08	Alosinae+Dorosomatinae, Clupeidae, Engraulinae	ENSDARG00000015866.8
gab1-2|2	GRB2-associated binding protein 1	OG0002631_1_1.inclade1.ortho2	6.17e-08, 6.17e-08, 6.17e-08	Alosinae+Dorosomatinae, Clupeidae, Engraulinae	ENSDARG00000037018.9
ppl	periplakin	OG0003698_1_1.inclade1.ortho1	7.74E-08	Alosinae+Dorosomatinae	ENSDARG00000101043.3
adgrg11-2|2	adhesion G protein-coupled receptor G11	OG0000992_1_1.inclade2.ortho1	1.08e-07, 1.08e-07, 1.20e-07	Alosinae+Dorosomatinae, Alosinae, *Alosa+Brevoortia*	ENSDARG00000041413.8
opn8b	opsin 8, group member b	OG0009183_2_1.unrooted-ortho	1.27e-07, 1.27e-07, 1.27e-07	Alosinae+Dorosomatinae, *Alosa+Brevoortia*, Alosinae	ENSDARG00000079045.5
si:ch211-207i1.2-3|7	si:ch211-207i1.2	OG0010770_1_1.inclade1.ortho1	2.06e-07, 2.06e-07, 2.06e-07, 2.06e-07	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	ENSDARG00000030107.10
unknown-555|965	NA	OG0015114_1_1.unrooted-ortho	3.17E-07	*Alosa+Brevoortia*	NA
katnbl1-1|2	katanin p80 subunit B-like 1	OG0005553_1_1.inclade1.ortho1	4.85E-07	Alosinae	ENSDARG00000042522.7
znf646-2|4	zinc finger protein 646	OG0007295_1_1.inclade1.ortho1	4.99E-07	Alosinae	ENSDARG00000061424.9
atp5pd	ATP synthase peripheral stalk subunit d	OG0009844_1_1.inclade1.ortho1	5.29E-07	Clupeidae	ENSDARG00000098355.2

Table 5 - Enriched GO terms

List with the thirteen enriched GO terms ranked by p-value. Enriched terms relate to mitochondrial functions and RNA processing. Figure 3 shows the relationships between these terms and the genes involved.

GO Identifier	GO Description	n. genes	P-value	Adj. p-value	Q-value	Orthogroups
GO:0005576	CC: extracellular region	42	2,77E-04	4,38E-02	4,29E-02	*ahsg1, apoa1a-1|2, apoa1a-2|2, apoeb-1|2, b2ml-1|2, bdnf-1|2, c1qtnf1, C20H6orf58, c4b, ccl20a.3, ccl34b.3, cfap53, chia.2-2|2, col10a1b, CR848035.1, cxcl11.1, cxcl32b.1-2|2, dcn, ecm1a-2|2, ecm1b, emilin1b, endouc, gdf3, hsd11b1la-2|2, hsp90ab1-2|2, ifi30, igf2b, igfbp2a, igfbp7, il10, il1b, kazald2, lipia, loxl3a, mgp, npc2, olfml3b, rnaset2, scg3, tg, thpo, timp2a-1|4*
GO:0005764	CC: lysosome	19	2,72E-05	1,08E-02	1,06E-02	*ap5s1, ctsc, ctsl.1-1|7, ctsl.1-6|7, cyb561a3a, fuca1.2, fuca2, glmp, ifi30, litaf, mfsd8, mios, npc2, rab9a-2|2, rnaset2, rnf152-2|2, si:dkey-228d14.5, vps41, znrf2b-1|2*
GO:0015986	BP: ATP synthesis coupled proton transport	5	2,93E-04	4,38E-02	4,29E-02	*atp5f1d, atp5l, atp5pd, atp5po, si:ch211-140m22.7-2|2*
GO:0005901	CC: caveola	6	2,02E-05	1,08E-02	1,06E-02	*cav1, cav3, cavin1b-1|3, cavin1b-2|3, cavin2b, zgc:172270*
GO:0006123	BP: mitochondrial electron transport, cytochrome c to oxygen	6	2,02E-05	1,08E-02	1,06E-02	*cox4i1l, COX5B, cox5b2, cox6a1, cox6a2, cycsb-1|2*
GO:0006412	BP: translation	23	4,87E-05	1,46E-02	1,43E-02	*eef1db, eif3hb, eif3i, eif3ja-1|3, eif5a-2|2, iars2, mrpl12, mrpl13, mrpl18, mrpl37, mrps18c, qars1, rfc1, rpl13, rpl14, rpl18, rpl23a, rps11, rps19, rps24-2|2, rps4x, rps6-1|2, tsfm*
GO:0005840	CC: ribosome	16	6,13E-05	1,46E-02	1,44E-02	*mrpl12, mrpl13, mrpl18, mrpl37, mrpl39-1|2, mrps18c, rfc1, rpl13, rpl14, rpl18, rpl23a, rps11, rps19, rps24-2|2, rps4x, rps6-1|2*
GO:0003735	MF: structural constituent of ribosome	16	2,06E-04	4,11E-02	4,03E-02	*mrpl12, mrpl13, mrpl18, mrpl37, mrps18c, mrps22, rfc1, rpl13, rpl14, rpl18, rpl23a, rps11, rps19, rps24-2|2, rps4x, rps6-1|2*

Table 6 - Enriched Reactome pathways

List of the thirteen enriched pathway compartments that are under selection in Clupeiformes. Most of them relate to mitochondrial functions and RNA processing. Figure 4 shows the relationships between these terms and the genes involved.

Reactome ID	Pathway Description	n. genes	P-value	Adj. p-value	Q-value	Orthogroups
R-DRE-163200	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	15	9,73E-07	7,25E-04	7,13E-04	*atp5f1d, atp5l, atp5pd, atp5po, etfdh, mt-nd1, mt-nd3, mt-nd6, ndufa11, ndufa7, ndufab1a, ndufb10, ndufb2, ndufb7, ndufv2*
R-DRE-1428517	The citric acid (TCA) cycle and respiratory electron transport	18	1,14E-04	2,13E-02	2,09E-02	*atp5f1d, atp5l, atp5pd, atp5po, dlst, etfdh, ldha, mt-nd1, mt-nd3, mt-nd6, ndufa11, ndufa7, ndufab1a, ndufb10, ndufb2, ndufb7, ndufv2, pdha1a*
R-DRE-72766	Translation	26	3,51E-06	1,31E-03	1,28E-03	*chchd1, eef1db, eef2b, eif3hb, eif3i, gadd45gip1, mrpl12, mrpl13, mrpl18, mrpl37, mrpl39-1|2, mrpl53, mrpl58, mrps18c, mrps22, mrps27, oxa1l, rpl13, rpl14, rpl18, rpl23a, rps11, rps19, rps24-2|2, rps4x, rps6-1|2*
R-DRE-5389840	Mitochondrial translation elongation	13	2,03E-04	2,49E-02	2,45E-02	*chchd1, gadd45gip1, mrpl12, mrpl13, mrpl18, mrpl37, mrpl39-1|2, mrpl53, mrpl58, mrps18c, mrps22, mrps27, oxa1l*
R-DRE-5419276	Mitochondrial translation termination	13	2,34E-04	2,49E-02	2,45E-02	*chchd1, gadd45gip1, mrpl12, mrpl13, mrpl18, mrpl37, mrpl39-1|2, mrpl53, mrpl58, mrps18c, mrps22, mrps27, oxa1l*
R-DRE-5368287	Mitochondrial translation	13	3,09E-04	2,88E-02	2,83E-02	*chchd1, gadd45gip1, mrpl12, mrpl13, mrpl18, mrpl37, mrpl39-1|2, mrpl53, mrpl58, mrps18c, mrps22, mrps27, oxa1l*
R-DRE-611105	Respiratory electron transport	11	5,43E-05	1,35E-02	1,33E-02	*etfdh, mt-nd1, mt-nd3, mt-nd6, ndufa11, ndufa7, ndufab1a, ndufb10, ndufb2, ndufb7, ndufv2*
R-DRE-6799198	Complex I biogenesis	10	1,55E-04	2,30E-02	2,26E-02	*mt-nd1, mt-nd3, mt-nd6, ndufa11, ndufa7, ndufab1a, ndufb10, ndufb2, ndufb7, ndufv2*

Table 7 - Enriched HGNC families

List of the five enriched HGNC families that are under selection in Clupeiformes. They relate to CD molecules, mitochondrial complexes I and V, apolipoproteins, and mitochondrial carriers (solute carrier family 25).

HGNC family	n. genes	P-value	Adj. p-value	Q-value	Orthogroups
CD molecules	34	9,82E-06	2,24E-03	2,13E-03	*alcama-1|2, alcama-2|2, asgrl2-1|2, bcam, cd28, cd44b, cd63-1|2, cd74a-2|2, cd81b, cdh2-1|2, crfb4, CU984600.1, entpd1, epcam, f11r.1-2|2, fut7-1|3, il13ra2, il6r-1|2, kel, mst1rb-1|2, rhag, sdc2-2|2, si:ch211-106j24.1-1|3, si:ch211-66e2.3-1|2, si:ch73-22o12.1-2|2, si:dkey-1c7.1-3|3, si:dkey-237i9.8, si:dkey-32n7.4-1|8, si:dkey-32n7.4-4|8, slc4a1a, tfr1a, tfr1b, tnfrsf1a, tnfsf13b*
Apolipoproteins	5	3,50E-04	1,99E-02	1,90E-02	*apoa1a-1|2, apoa1a-2|2, apoda.2, apoeb-1|2, apooa*
Mitochondrial complex V: ATP synthase subunits	5	1,50E-04	1,71E-02	1,63E-02	*atp5f1d, atp5l, atp5pd, atp5po, si:ch211-140m22.7-2|2*
NADH:ubiquinone oxidoreductase supernumerary subunits	6	9,67E-04	4,41E-02	4,19E-02	*ndufa11, ndufa7, ndufab1a, ndufb10, ndufb2, ndufb7*
solute carrier family 25	8	2,47E-04	1,88E-02	1,78E-02	*slc25a12-1|2, slc25a12-2|2, slc25a19, slc25a20, slc25a26, slc25a28, slc25a38b, slc25a47a-1|2*

Table 8 - Genes Under Recurrent Selection

List with the genes that were selected in at least four branches at the same time.

Orthogroup Symbol	Gene Description	Phylogroups	Adjusted p-values	Ensembl ID
CR846080.1-2|2	None	Alosinae+Dorosomatinae, *Alosa+Brevoortia*, Alosinae, Clupeidae, Engraulinae	1.85e-03, 1.85e-03, 1.85e-03, 1.85e-03, 1.85e-03	ENSDARG00000076211.4
si:ch211-15j1.5	si:ch211-15j1.5	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae, Sardinae	2.86e-03, 2.86e-03, 2.86e-03, 2.86e-03, 2.86e-03	ENSDARG00000091912.4
si:ch211-207i1.2-3|7	si:ch211-207i1.2	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	2.06e-07, 2.06e-07, 2.06e-07, 2.06e-07	ENSDARG00000030107.10
mylipa-1|3	myosin regulatory light chain interacting protein a	Alosinae, Clupeidae, Engraulinae, Alosinae+Dorosomatinae	8.25e-06, 8.25e-06, 8.25e-06, 1.30e-05	ENSDARG00000008859.8
apoa1a-1|2	apolipoprotein A-Ia	Alosinae, Engraulinae, Alosinae+Dorosomatinae, Clupeidae	1.11e-05, 5.25e-03, 1.36e-02, 1.36e-02	ENSDARG00000012076.9
numa1-2|2	nuclear mitotic apparatus protein 1	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	5.27e-05, 5.27e-05, 5.27e-05, 5.27e-05	ENSDARG00000102483.2
si:dkey-237j10.2-2|2	si:dkey-237j10.2	Alosinae+Dorosomatinae, Alosinae, Engraulinae, Clupeidae	1.56e-04, 1.92e-04, 1.26e-02, 1.45e-02	ENSDARG00000094025.3
unknown-565|965	NA	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	1.80e-04, 1.80e-04, 1.80e-04, 1.95e-04	NA
si:cabz01007794.1	si:cabz01007794.1	Alosinae, Engraulinae, Clupeidae, Alosinae+Dorosomatinae	2.10e-04, 1.42e-03, 1.47e-03, 3.20e-03	ENSDARG00000105590.2
si:dkeyp-121d4.3-3|6	si:dkeyp-121d4.3	Alosinae, Alosinae+Dorosomatinae, Clupeidae, Engraulinae	7.18e-04, 7.21e-04, 7.21e-04, 7.21e-04	ENSDARG00000089355.3
s100w	S100 calcium binding protein W	Alosinae, Alosinae+Dorosomatinae, Clupeidae, Engraulinae	8.26e-04, 2.74e-02, 2.74e-02, 2.74e-02	ENSDARG00000101181.2
epb41b	erythrocyte membrane protein band 4.1b	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	2.31e-03, 2.31e-03, 2.31e-03, 2.31e-03	ENSDARG00000029019.6
zp3a.2-1|4	zona pellucida glycoprotein 3a, tandem duplicate 2	Alosinae+Dorosomatinae, Clupeinae, Clupeidae, Engraulinae	2.52e-03, 1.09e-02, 2.48e-02, 3.04e-02	ENSDARG00000042130.6
no-symbol-5|9	APC down-regulated 1	*Alosa+Brevoortia*, Alosinae+Dorosomatinae, Alosinae, Clupeidae	2.98e-03, 3.93e-02, 3.93e-02, 3.93e-02	ENSDARG00000098203.2
unknown-422|965	NA	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	5.17e-03, 5.17e-03, 5.17e-03, 5.17e-03	NA
eif4g2a-2|2	eukaryotic translation initiation factor 4, gamma 2a	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	8.22e-03, 8.22e-03, 8.22e-03, 8.22e-03	ENSDARG00000020377.10
si:dkey-32n7.4-4|8	si:dkey-32n7.4	Alosinae+Dorosomatinae, Alosinae, *Alosa+Brevoortia*, Engraulinae	2.03e-02, 2.03e-02, 2.03e-02, 2.03e-02	ENSDARG00000002956.12
unknown-185|965	NA	Alosinae+Dorosomatinae, Alosinae, Clupeidae, Engraulinae	3.42e-02, 3.42e-02, 3.42e-02, 3.42e-02	NA

Figures

Figure 1 - Phylogeny

Consensus phylogeny inferred by ExaBayes and four-fold degenerate sites. Branch length represents the number of observed mutations. All branches obtained 100% bootstrap support. Branches marked in red are the ones tested for selection in the branch-site test.

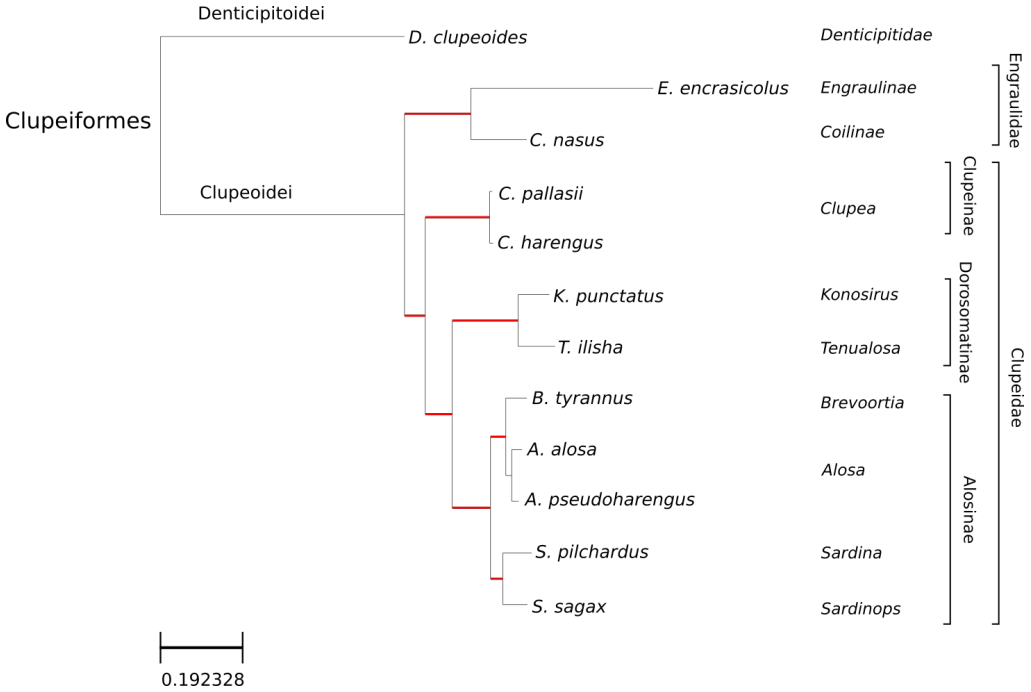


Figure 2 - Busco

MultiQC report from BUSCO results based on the Single Copy Orthologs in the *Actinopterygii* dataset. Transcriptomes derived from assembled genomes obtained the highest results (*D. clupeoides* and *C. harengus*), followed by *S. pilchardus* and *A. pseudoharengus*. The set of the *Actinopterygii* Single Copy Orthologs consists of 4,584 sequences.

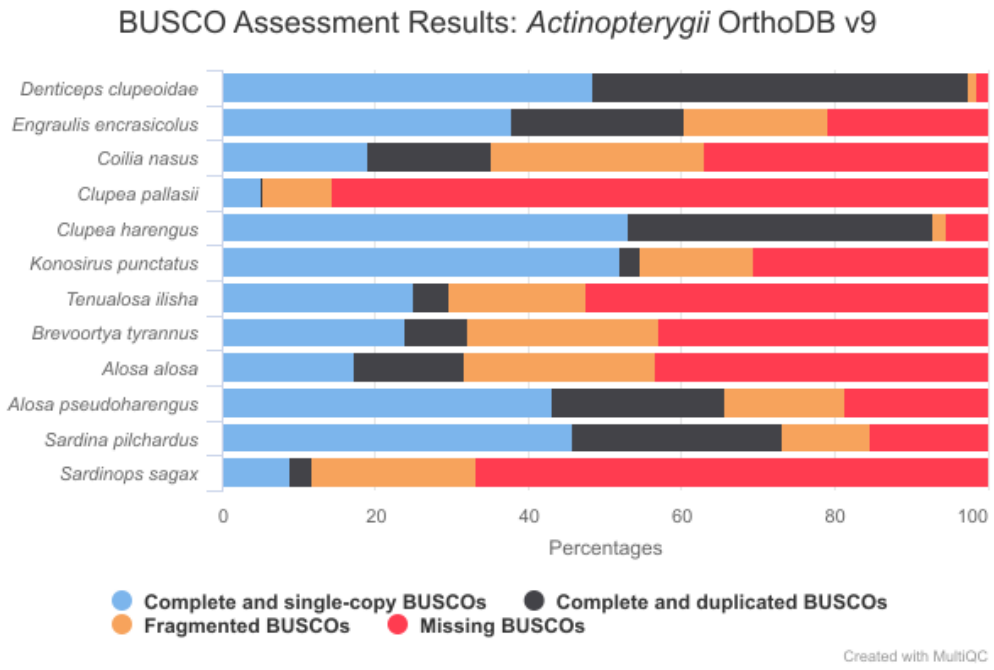


Figure 3 - GO Concept Network

GO/Orthogroup concept network. Representation of the enriched GO terms derived from Zebrafish and the orthogroups associated. We observe three superclusters: extracellular proteins, lysosome, and the ribosome, and three smaller ones: mitochondrial complexes IV and V, and caveolae.

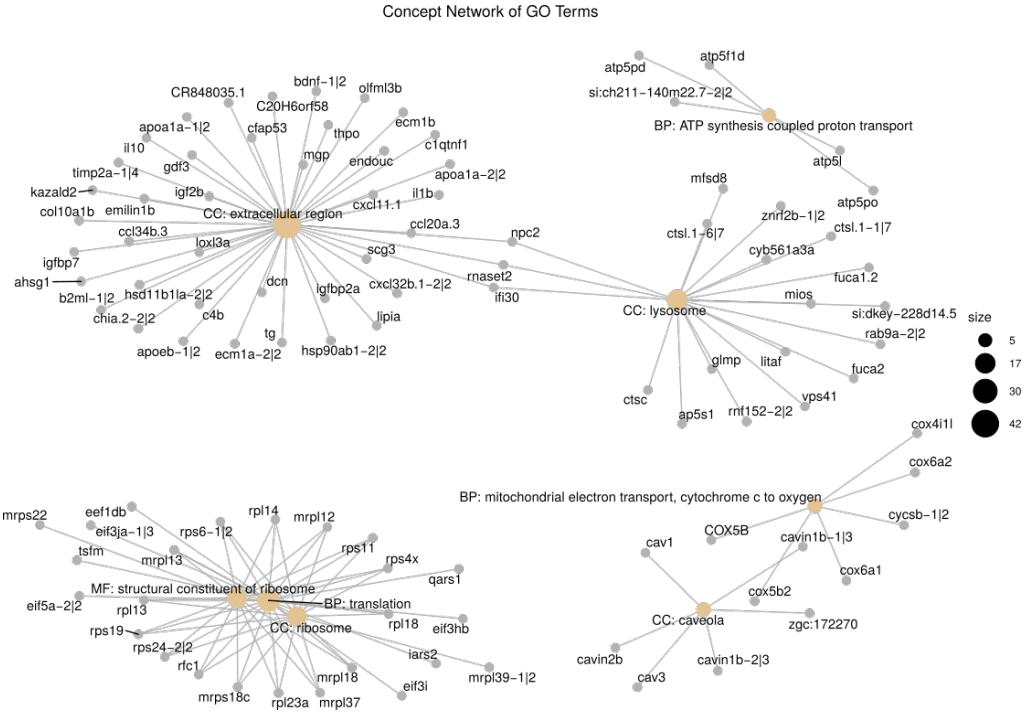


Figure 4 - Reactome Concept Network

Reactome pathways / Orthogroup concept network. Representation of the enriched Reactome pathways derived from the annotations for Zebrafish and the orthogroups associated. We observe two clusters: one associated with mitochondria, a second one on RNA translation.

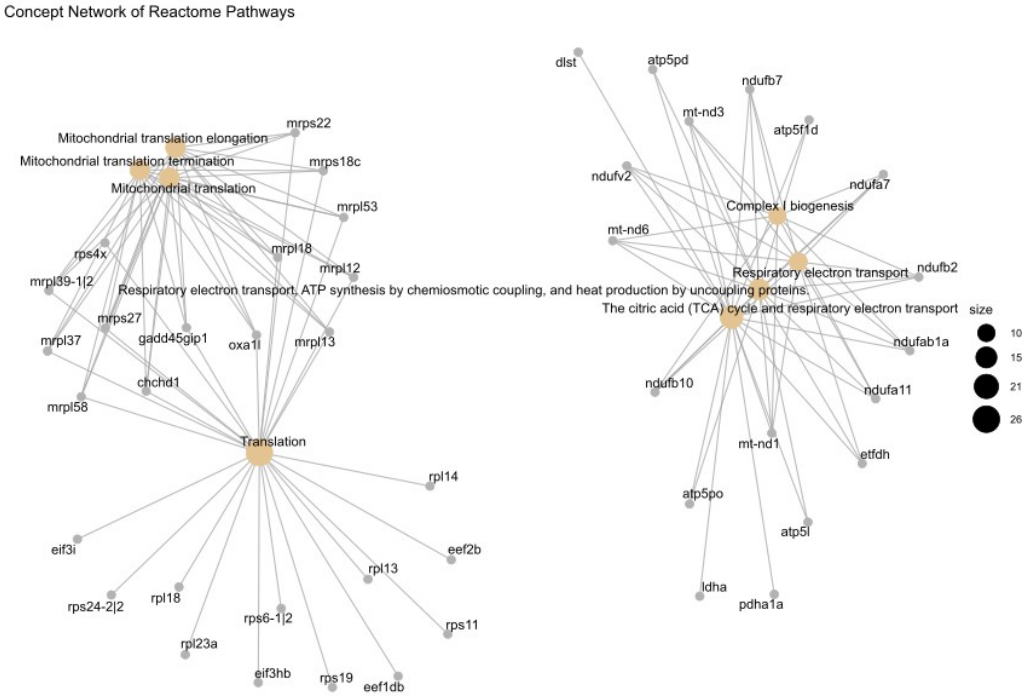


Figure 5 - HGNC concept network

HGNC family / Orthogroup concept network. Representation of the enriched gene families derived from the human HGNG annotations inferred for the Clupeiformes orthogroups. We observe five clusters: CD molecules, complexes I and V of the Electron Transport Chain, apolipoproteins, and mitochondrial transporters.

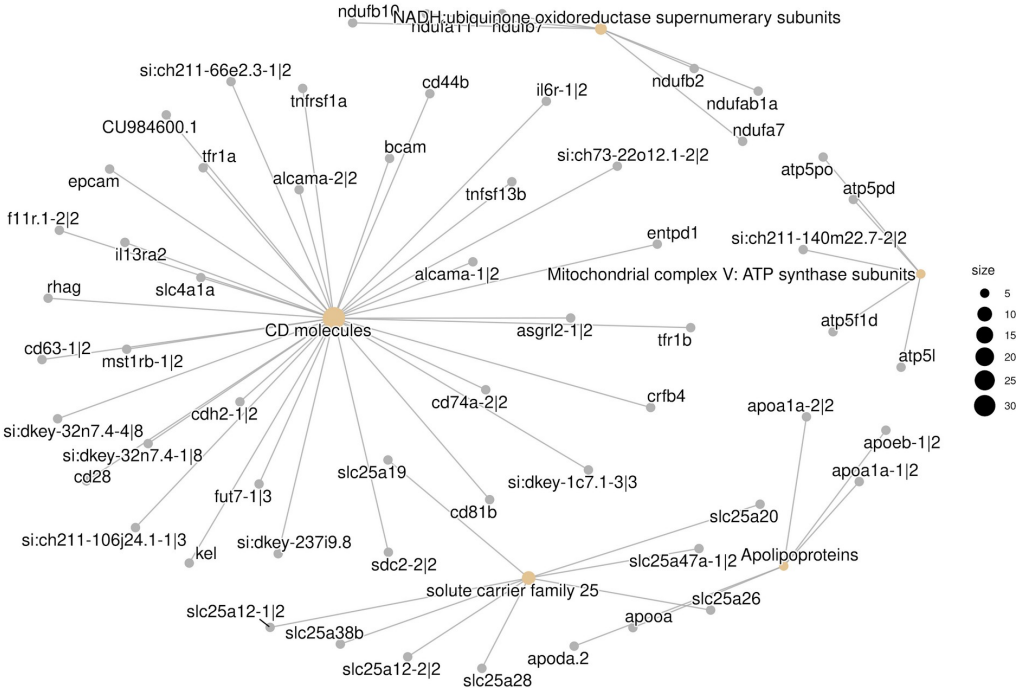
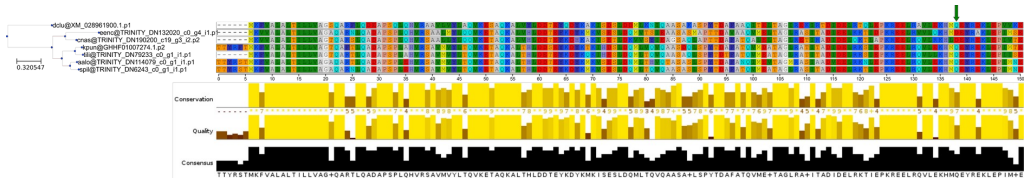


Figure 6 - Alignment of *apoa1a-1|2*

First 151 aligned positions of the protein *apoa1a-1|2*. Branches in red in the phylogenetic tree denote that were under selection after p-value correction. The arrow in green shows the only site under positive selection for the family *Engraulidae.* Additional branches under selection were *Alosinae and Alosinae+Dorosomatinae* (and *Clupeidae* due to the omission of *Clupea* proteins).



Supplementary Tables

ST01- Transcriptome and RNA-Seq dataset accession numbers

ST02 - Possible duplicated genes and orthogroups without an annotation

ST03 - Selected genes

ST04 - Genes Under Recurrent Selection