INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR

**The Evolution of Fatty Acid Metabolism in Chordates**

Mónica Lopes Marques

U.PORTO

**Mónica Lopes Marques.** The Evolution of Fatty Acid Metabolism in Chordates

**D.**ICBAS **2017**

U.PORTO

DOUTORAMENTO EM CIÊNCIAS BIOMÉDICAS

# The Evolution of Fatty Acid Metabolism in Chordates

Mónica Lopes Marques

**D**

**2017**

MÓNICA LOPES MARQUES

# The Evolution of Fatty Acid Metabolism in Chordates

Tese de Candidatura ao grau de Doutor em Ciências Biomédicas submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Orientador – Luís Filipe Costa Castro
Categoria – Investigador Auxiliar
Afiliação – Centro Interdisciplinar de Investigação Marinha e Ambiental

Coorientador - Miguel Alberto Fernandes Machado e Santos
Categoria – Professor Auxiliar
Afiliação – Faculdade de Ciências da Universidade do Porto

Coorientador - Eduardo Jorge Sousa da Rocha
Categoria – Professor Catedrático
Afiliação – Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

## TABLE OF CONTENTS

## AUTHOR STATEMENT

This thesis is organized into nine chapters. Chapter I and Chapter II consist of a general introduction intended to give an all-purpose background and supporting information on evolutionary mechanisms and FA metabolic pathways covered in detail in the remaining Chapters. Chapter III to Chapter VI correspond to several projects developed during the doctoral program presented here as independent articles. Finally, Chapter VII and VIII contain a general discussion and coalescing of the main findings of the presented work and final remarks, culminating with future perspectives, Chapter IX.

The work included in this thesis was totally or partially executed by the candidate, in close cooperation/co-authorships with supervisors and other researchers. In detail, the conception and full draft of Chapter I, Chapter II, Chapter VII, Chapter VIII and Chapter IX are the sole responsibility of the candidate, with revisions from the supervising team. In the remaining chapters the candidate made substantial contributions. The articles included herein will not appear in other theses or dissertations. During the PhD work program, the candidate actively participated in other research projects, which entailed additional publications not included in this thesis (check curriculum vitae).

In summary, this thesis includes seven articles published in peer reviewed international journals, one article under review and one article in final preparation for submission. All articles were integrated within the thesis as chapters (first author or joint first authors are underlined):

**Chapter III- FA Activation**

<u>Castro, L. F. C., M. Lopes-Marques</u>, J. M. Wilson, E. Rocha, M. A. Reis-Henriques, M. M. Santos and I. Cunha (2012). "A novel Acetyl-CoA synthetase short-chain subfamily member 1 (*Acss1*) gene indicates a dynamic history of paralogue retention and loss in vertebrates." *Gene* **497**(2): 249-255.

<u>Lopes-Marques</u>, M., I. Cunha, M. A. Reis-Henriques, M. M. Santos and L. F. C. Castro (2013). "Diversity and history of the long-chain acyl-CoA synthetase (*Acsl*) gene family in vertebrates." *BMC Evolutionary Biology* **13**(1): 271.

**Chapter IV- FA Biosynthesis**

Monroig, Ó., M. Lopes-Marques, J. C. Navarro, F. Hontoria, R. Ruivo, M. M. Santos, B. Venkatesh, D. R. Tocher and L. F. C. Castro (2016). "Evolutionary functional elaboration of the *Elovl2/5* gene family in chordates." *Scientific Reports* **6**: 20510.

Lopes-Marques, M., R. Ozório, R. Amaral, D. R. Tocher, Ó. Monroig and L. F. C. Castro (2016). "Molecular and functional characterization of a *fads2* orthologue in the Amazonian teleost, *Arapaima gigas*." *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **203**: 84-91.

Lopes-Marques, M., N. Kabeya, Y. Qian, R. Ruivo, M. Santos, B. Venkatesh, L. F. C. Castro and O. Monroig (2017). "Gene duplication and loss underscore the vertebrate efficiency in completing long-chain polyunsaturated fatty acids biosynthesis." IN PREPARATION.

**Chapter V- FA β-Oxidation**

Lopes-Marques, M., I. L. S. Delgado, R. Ruivo, Y. Torres, S. B. Sainath, E. Rocha, I. Cunha, M. M. Santos and L. F. C. Castro (2015). "The Origin and Diversity of Cpt1 Genes in Vertebrate Species." *PLoS One* **10**(9): e0138447.

Lopes-Marques, M., R. Ruivo, I. Delgado, J. M. Wilson, N. Aluru and L. F. C. Castro (2015). "Basal Gnathostomes Provide Unique Insights into the Evolution of Vitamin B12 Binders." *Genome Biology and Evolution* **7**(2): 457-464.

**Chapter VI- Protein Digestion and Gastric Proteases**

Castro, L. F. C., M. Lopes-Marques, O. Gonçalves and J. M. Wilson (2012). "The Evolution of Pepsinogen C Genes in Vertebrates: Duplication, Loss and Functional Diversification." *PLoS One* **7**(3): e32852.

Lopes-Marques, M., R. Ruivo, E. Fonseca, A. Teixeira and L. F. C. Castro (2017). "Unusual loss of chymosin in mammalian lineages parallels neonatal immune transfer strategies." UNDER REVIEW.

## ACKNOWLEDGMENTS

First, I must express my sincere utmost gratitude to my advisor Filipe Castro, for the opportunity to join his research team, where I have never stopped learning and growing as researcher from day one. Thank-you for always pushing me a bit further, sharing your immense and humbling knowledge, for your guidance, patience, support, and for the inspiring discussions with piercing ideas that lead to many sleepless nights. I could not have asked for a better advisor and mentor for my PhD.

I would also like thank Prof. Miguel Santos for his guidance, support, patience and for organizing the group meetings with stimulating discussions, broadening my research interests and bringing novel perspectives. Next, I would like to thank Prof. Eduardo Rocha for welcoming me to his research team and laboratory, for his encouragement, support and patience, which I very much appreciated.

I am also indebted to CIIMAR – Interdisciplinary Centre of Marine and Environmental Research and ICBAS – Abel Salazar Institute of Biomedical Sciences, for making me feel at home and providing the infrastructure and funding that made this work possible.

I am very grateful to Prof. Óscar Monroig at the Institute of Aquaculture University of Stirling for welcoming me to his laboratory, giving me the opportunity to learn aside him, for his patience and support in drafting the corresponding manuscripts.

I would like to thank Catarina Cruzeiro an old and dear friend. I am indebted for your patience, support and friendship, and may our adventures, together, lead us to bright futures. I am also very grateful to Raquel Ruivo for her friendship, sharing her overwhelming knowledge, critical assessments and for being the perfect *partner in crime*. Next, Maria João and Ana Capitão my partners in this PhD voyage thank-you for your honest friendship, encouragement, team work, spontaneity, positivity, and for making this PhD an inspiring joyful journey. I would like to thank Elza Fonseca for her audacity, support and hard work she put into helping me with several projects.

I would like to acknowledge all the previous, current and new members of CIIMAR who have crossed my path and that in one way or another helped me along this journey,

Elsa Froufe, Jonathan Wilson, Tânia Madureira, Alice Ramos, Ivone Pinheiro, Odete Gonçalves, Miguel Cordeiro, Fernanda Malhão, Ana André, Inês Pascoa, Inês Delgado, Diogo Martins, Diogo Oliveira, Ana Teixeira, Susana Barbosa, and André Machado.

I am also very grateful to Luísa Azevedo for introducing me to evolutionary biology for her support and friendship.

For making my life a wonderful journey I must thank my friends Olga Pires, Hugo Pereira, Mª João Peixoto, Ana Capitão, Raquel Ruivo, Inês Pascoa, Rita Santos, Estefânia Martins, Balázs Gárgyán, Mª João Machado, Andreia Irina, Carmo Cardoso, Sérgio Mendes, Ricardo Santos, Liliana Gonçalves, Luís e Rute Sobral, they say that friends are the family you choose and I am so lucky to have such a great friend family.

I am exceptionally indebted to Sérgio Fernandes, for being my partner in this journey called life, for being my best friend, for making me a better person and for supporting me unconditionally, Thank-you!

Finally, I would like to express my profound gratitude to my family my parents, who are my pillars and to my brother and his family, I have no words that describe all I that owe you.

I would like to acknowledge the European Society for Comparative Endocrinology (ESCE), for the best poster prize awarded at the 26th Conference of European Comparative Endocrinologists (CECE) 2012, which I used to support my accommodation and registration in the Congress of the European Society for Evolutionary Biology (ESEB) in 2013.

I am also grateful to EMBL- European Bioinformatics Institute for awarding me a travel grant and accommodation that provided me the opportunity to attend the EMBO-Practical Course on Analysis of High-Throughput Sequencing Data held in October 2014 at the EMBL-ENI, Wellcome Trust Genome Campus, Hinxton UK.

I would like to acknowledge FCT – Fundação da Ciência e Tecnologia for supporting me with a PhD grant SFRH/BD/84238/2012. This work was further supported by the ICBAS, via the PhD Program in Biomedical Sciences, and by the research line INSEAFOOD within the INNOVMAR project (NORTE-01-0145-FEDER-000035), North Portugal Regional Operational Program (NORTE 2020) under the Portugal 2020 Partnership agreement through the European Regional development fund (ERDF):

# roll the dice

if you're going to try, go all the
way.
otherwise, don't even start.

if you're going to try, go all the
way.
this could mean losing girlfriends,
wives, relatives, jobs and
maybe your mind.

go all the way.
it could mean not eating for 3 or 4 days.
it could mean freezing on a
park bench.
it could mean jail,
it could mean derision,
mockery,
isolation.
isolation is the gift,
all the others are a test of your
endurance, of
how much you really want to
do it.
and you'll do it
despite rejection and the worst odds
and it will be better than
anything else
you can imagine.

if you're going to try,
go all the way.
there is no other feeling like
that.
you will be alone with the gods
and the nights will flame with
fire.

do it, do it, do it.
do it.

all the way
all the way.

you will ride life straight to
perfect laughter, its
the only good fight
there is.

- Charles Bukowski

## LIST OF ABBREVIATIONS

| | |
|---|---|
| 2R WGD | Two rounds of whole genome duplication |
| 3R WGD | Teleost specific genome duplication or three rounds of whole genome duplication |
| 4R WGD | Four rounds of whole genome duplication |
| ABCD | ATP Binding Cassette Subfamily D |
| ACAC | Acetyl-CoA Carboxylase Alpha |
| ACOT | Acyl-CoA Thioesterase |
| ACOX | Acyl-CoA Oxidase |
| ACSBG | Bubble gum Acyl-Coenzyme A synthetase |
| ACSL | Long chain Acyl-Coenzyme A synthetase |
| ACSS | Short chain Acyl-Coenzyme A synthetase |
| ALA | α-linolenic acid |
| ARA | Arachidonic acid |
| ATP | Adenosine triphosphate |
| BC | Before Crist |
| CCR5 | C-C chemokine receptor type 5 |
| CoA | Coenzyme A |
| CPT | Carnitine Palmitoyltransferase |
| CRAT | Carnitine O-Acetyltransferase |
| CROT | Carnitine O-Octanoyltransferase |
| DGAT | Diacylglycerol O-Acyltransferase |
| DHA | Docosahexaenoic acid |
| DNA | Deoxyribonucleic acid |
| EFA | Essential Fatty acid |
| ELOVL | Fatty Acid Elongase |
| EPA | Eicosapentaenoic acid |
| FA | Fatty Acid |
| FADS | Fatty Acid desaturase |
| FASN | Fatty Acid Synthase |
| Fe | Iron |
| GIF | Gastric intrinsic factor |
| GPAM | Mitochondrial Glycerol-3-Phosphate Acyltransferase |

| | |
|---|---|
| HIV | Human immunodeficiency virus |
| HOX | Homeobox genes |
| LCAT | Lecithin-Cholesterol Acyltransferase |
| LC-PUFA | Long chain polyunsaturated fatty acid |
| LIPA, LIPE | Lipase A, Lipase E |
| LIPC, LIPG, LIPH | Lipase C, Lipase G, Lipase H |
| LOA | Linoleic acid |
| LPL, | Lipoprotein Lipase |
| MC-PUFAS | Medium chain polyunsaturated fatty acid |
| MOGAT | Monoacylglycerol O-Acyltransferase |
| MYA | Million years ago |
| MYH16 | Myosin heavy chain 16 |
| PGA, PGB, PGC | Pepsinogen A, Pepsinogen B, Pepsinogen C |
| PGF, CYM | Pepsinogen F, Chymosin |
| PlA1a | Phospholipase A1 |
| PPAR | Peroxisome proliferator-activated receptor |
| PUFA | Poly unsaturated fatty acid |
| SCD | Stearoyl-CoA Desaturase |
| SOAT | Sterol O-Acyltransferase |
| TCN 2 | Transcobalamin 2 |
| TCN1 | Transcobalamin 1, or haptocorrin |

## ABSTRACT

Lipids are vital components of all living organisms, and together with proteins and carbohydrates constitute a major building block of life. Lipids are involved in numerous biological functions, contributing for organism homeostasis by supplying and storing energy. They also have a regulatory role serving as endogenous ligands or signaling molecules and play an important structural role being a considerable part of bio-membranes. According to their chemical structure they can be grouped into several categories such as: fatty acids (FA), glycerolipids, glycerophospholipids, sphingolipids, prenol lipids, saccharolipids, and polyketides. Among those, FAs are particularly relevant since the majority of complex lipids are obtained from the elaboration of FA and therefore constitute a considerable portion of total lipid fraction. Although FA composition and metabolism is known to vary among vertebrate species, in many cases an integrated evolutionary view of the metabolic FA pathways and genetic repertoire is yet to be produced. In particular, the importance and potential impact of genomic processes such as: the 2 rounds of whole genome duplication (2R WGD) that occurred in the invertebrate/vertebrate transition, the teleost specific genome duplication (3R WGD), tandem gene duplications, gene loss and mutation, which have been largely overlooked. Thus, to clarify the underpinning of the observed distinct FA compositions and metabolic capacities in vertebrates it is necessary to investigate the genetic machinery involved in these metabolic pathways across several vertebrate lineages. Additionally, it is essential to link the evolutionary life trajectories in vertebrates such as: colonization of marine *vs* freshwater ecosystems, access to novel or alternative dietary sources or the colonization of terrestrial habitats, to the genetic repertoire and metabolic capability displayed by different species.

In this context, three rate limiting FA metabolic pathways were investigated: FA activation; FA β-oxidation; and FA biosynthesis. In the FA activation pathway, the invertebrate/vertebrate transition entailed an expansion of the long chain acyl-coenzyme A synthetase (*ACSL*) gene family coincident with the 2R WGD. Additionally, uncharacterized paralogues from the *ACSL* and short chain acyl-coenzyme A synthetase (*ACSS*) gene families were uncovered in the several vertebrate lineages, displaying a dynamic history of differential paralogue retention. Moreover, the teleost 3R WGD also

contributed for the expansion of the genetic machinery involved in FA activation revealing an elaboration of these pathways in the teleost lineage.

On the other hand, FA biosynthesis, specifically the long chain polyunsaturated FA (LC-PUFA) biosynthesis pathway, is known to be impaired in some teleost species. Here, I investigated the FA elongase and FA desaturase gene families (*Elovl* and *Fads*) to reveal that the 2R WGD significantly contributed for the functional elaboration of the *Elovl* gene family in vertebrates, while tandem gene duplication spawned *Fads* diversification in the vertebrate ancestor. The reexamination of these gene families revealed an unforeseen *fads1* orthologue in *Lepisosteus oculatus*, *Polypterus senegalus*, *Anguilla anguilla*, indicating that the loss of *fads1* occurred after the divergence of basal teleost lineages (holostei, polypteriformes, elopomorpha) and clarifying the evolutionary history of this gene family. Next, functional characterization of *elovl* the invertebrate amphioxus as well as, *elovl* and *fads* from agnathans, basal gnathostomes and teleosts suggests that the acquisition of the full LC-PUFA biosynthetic pathway took place in the ancestor of the gnathostomes, confirming that impairment in LC-PUFA biosynthesis observed in many teleost species is due to secondary gene loss.

Concerning the β-Oxidation pathway the investigation of the Carnitine palmitoyltransferase 1 (*CPT1*) gene family revealed again an evolutionary history with differential paralogue retention in several lineages, and retention of 3R WGD duplicates in teleosts. Regarding the B12 binder gene family, reexamination of the evolutionary history revealed that the initial expansion of this gene family took place with the 2R WGD, which was followed by 2 events of gene loss, one in the ancestor of sarcopterygii and actinopterygii and the second in the teleost lineage. Additionally, gene expansion by tandem duplication is observed in basal tetrapods paralleling the transition to terrestrial habitats and access of novel dietary sources.

Similarly, protein metabolism as well as digestive protein processes, have also been significantly impacted by gene/genome duplication and gene loss. The investigation of the pepsinogen C (*PgC*) gene family revealed a larger genetic repertoire than anticipated and that this expansion again was coincident with the colonization of terrestrial habitats and access to novel food sources. Interestingly, an alternative

evolutionary history is found for the neonatal protease chymosin (*Cmy*). Although, it was previously shown that *Cmy* is a pseudogene in humans, I described an unprecedented number of independent gene loss events in various mammalian lineages, suggesting a correlation to alternative immune transfer strategies in neonatals.

The investigation of the evolutionary history of several gene families directly involved in lipid and protein metabolism revealed the impact of genomic processes such as duplication (2R WGD, 3R WGD), gene loss and mutation in the elaboration of several metabolic pathways in vertebrates, thus contributing towards vertebrate diversification. The findings reported here illustrate the power of comparative genomics in the Genome Era and provide important clues well beyond the field of evolutionary biology, with significant impacts in fields such as animal nutrition and aquaculture.

## RESUMO

Os lípidos são componentes vitais em todos os seres vivos. Em conjunto com as proteínas e os hidratos de carbono, integram as unidades fundamentais das quais todos os organismos são constituídos. Os lípidos executam várias funções biológicas; contribuem para a homeostase através da produção de energia ou armazenamento, podem ter funções de regulação sendo ligandos endógenos ou moléculas sinalizadoras e cumprem uma função estrutural, sendo o maior constituinte de biomembranas. De acordo com a sua estrutura química, os lípidos podem ser classificados em vários grupos: ácidos gordos (AG), glicerolípidos, glicerofosfolípidos, esfingolípidos, lípidos prenólicos, sacarolípidos e policetídeos. Dos quais os AG são de particular relevância, visto que são a unidade básica a partir do qual os lípidos mais complexos são elaborados, constituindo assim uma fracção considerável dos lípidos totais. Apesar de ser conhecido que a composição em AG e as vias metabólicas dos AG variam entre os vertebrados, é necessário elaborar uma perspectiva evolutiva sobre o reportório de genes envolvidos nestas vias metabólicas. Neste contexto, o impacto dos processos genómicos como as duas rondas de duplicação do genoma que ocorreram na transição invertebrados/vertebrados (2R WGD), a duplicação específica dos teleósteos (3R WGD), duplicações de genes em tandem, perda de genes e mutações tem sido largamente ignorados. Assim, para clarificar os processos subjacentes às diferenças observadas na composição em AG e respectivas vias metabólicas nos vertebrados é necessário catalogar a maquinaria genética interveniente nestas vias em várias linhagens de vertebrados. Por fim, é essencial integrar o reportório genético e aptidões metabólicas das várias linhagens de vertebrados aos correspondentes percursos evolutivos, tais como: colonização de ecossistemas marinhos ou de água doce, acesso a novos recursos alimentares, e colonização de ambientes terrestres.

Neste contexto, foram seleccionadas para investigação 3 vias limitantes do metabolismo dos AG; ativação dos AG, biossíntese de AG e β-oxidação de AG, sendo investigada a história evolutiva de um conjunto de genes intervenientes nestas vias metabólicas. Na via de ativação dos AG verificou-se que a 2R WGD levou a uma expansão da família de genes acyl-coenzima A sintetase de cadeia longa (*ACSL*). Foram ainda encontrados novos membros *ACSL* e das acyl-coenzima A sintetase de cadeia

curta (*ACSS*) na linhagem dos teleósteos resultantes do 3R WGD. Revelou-se assim uma história dinâmica de retenção diferencial de genes parálogos e o contributo do 3R WGD para a expansão de genes das vias de activação de AG na linhagem dos teleósteos.

Por outro lado, a via de biossíntese de AG polinsaturados de cadeia longa (LC-PUFAS) encontra-se interrompida na linhagem dos teleósteos. Aqui, a investigação dos genes que codificam as elongases de AG e as desaturases de AG (*Elovl* e *Fads*) revelou que o 2R WGD contribuiu para a elaboração funcional das *elovls*, ao passo que a esta elaboração nas *fads* se deve uma duplicação em tandem na base dos vertebrados. A reanálise desta família de genes revela uma inesperada retenção de um ortólogo *fads1* em teleósteos (*Anguilla anguilla, Lepisosteus oculatus*, *Polypterus senegalus*) indicando que a sua perda ocorreu após a divergência destas linhagens e clarificando a história evolutiva destes genes. De seguida, a caracterização funcional das *elovl* no anfioxo e das *elovl* e *fads* nos ágnatos, gnatostomas basais e teleósteos revela que a aquisição da via completa de síntese de LC-PUFAs ocorreu nos gnatostomas basais, confirmado que as limitações desta via observadas em teleósteos se devem a perdas posteriores.

Relativamente às famílias génicas envolvidas na β-oxidação, observa-se novamente a retenção diferencial de parálogos da carnitina palmitoiltransferase 1 (*CPT1*) em várias linhagens de vertebrados e com a retenção adicional de parálogos resultantes do 3R WGD nos teleósteos. A reavaliação da história evolutiva da família génica dos transportadores da B12 revela uma história alternativa onde a expansão inicial se deu com o 2R WGD seguindo-se dois eventos de perda, um no ancestral sarcopterígios e dos actinopterígios e o segundo na linhagem dos teleósteos. Também se observa uma expansão desta família na base dos tetrápodes, coincidente com a colonização de habitats terrestres e a acesso a novas fontes alimentares.

À semelhança do metabolismo dos AGs, o metabolismo proteico, nomeadamente da família de genes das protéases gástricas, também foi afetado por eventos de duplicação de genoma e/ou de genes, perda de genes e mutação. A análise da família de genes do pepsinogénio C (*PgC*) revela que esta família retém um reportório génico maior do que o antecipado. Curiosamente, também esta expansão coincide com a transição dos vertebrados para habitats terrestres com o acesso a novas fontes dietéticas. Comparativamente, a protéase neonatal quimosina (*Cmy*) revela uma história evolutiva alternativa, apesar de a *Cmy* ser um conhecido pseudogene em humanos. Na análise desta família encontrou-se um número inédito de eventos independentes de pseudogenização em várias linhagens de mamíferos, sugerindo uma correlação com mecanismos alternativos de transferência de imunidade em neonatais.

A investigação de várias famílias génicas envolvidas no metabolismo dos lípidos e das proteínas revela o impacto dos processos genómicos tais como 2R WGD e 3R WGD, duplicação génica, perda génica e mutação na elaboração das vias metabólicas em vertebrados contribuindo para a sua diversidade. Os resultados reportados aqui ilustram o poder da genética comparativa na presente Era genómica e providenciam indicações importantes com impacto em áreas além da biologia evolutiva, tais como a nutrição animal e aquacultura.

# CHAPTER I

## INTRODUCTION

# CHAPTER I – INTRODUCTION

## I.1 HISTORICAL PERSPECTIVE ON EVOLUTION AND GENETICS

Evolution has been a fascinating subject for many years with the first proto-evolutionary ideas appearing as early as 550 BC by the Greek philosopher Anaximander of Miletus (Kocandrle *et al.*, 2013; Trevisanato, 2016). The most striking ideas postulated by Anaximander were that the first biological systems, or beings, emerged in an aquatic environment that he referred as "*from the moist*", and that humans developed from an animal that resembled a fish (Kocandrle *et al.*, 2013; Trevisanato, 2016). Nevertheless, in the 24 centuries that separate Anaximander of Miletus from Darwin several other significant contributions were made in the evolutionary biology field.

In 1735, Carl Linnaeus publishes the *Systema Naturae*, organizing species according to complexity and naming each species with a binomial nomenclature, which is still used today (Linnaeus, 1735). This was followed by the publication of the *"Essay on the principle of populations"* in 1798 by Thomas Robert Malthus. This essay inspired both Darwin and Alfred Wallace in the development of the theory of natural selection, based on the model presented by Malthus that continued population growth would outgrow resources (Richards, 2009; Ruse, 2009). The publication of the gradualism theory by James Hutton in 1795 introduced the notion of geologic time which was followed by the publication of the principal of uniformitarianism in 1830 by Charles Lyell, which postulated that geological process operating at the beginning of time are the same as those observed today (Darwin, 1887). The ideas postulated by James Hutton and Charles Lyell are latter reflected in evolutionary theory presented by Darwin. Yet, meanwhile, the first evolutionary theory appears in 1809, published by a French naturalist Jean Baptiste Lamarck, who advocated that when the environment changes animals would also change to better adapt; these alterations occurred through the use or disuse of certain characteristics/features and that the acquired characteristics could be transmitted to the offspring (Richards, 2009).

Finally, in 1858, twenty-two years after completing his voyage on the HMS Beagle, Darwin is spurred to complete his essay on evolution, after receiving a letter from Alfred Wallace, as he later acknowledged in his autobiographic letters:

"*...But my plans were overthrown, for early in the summer of 1858 Mr. Wallace, who was then in Malay archipelago, sent me an essay On the Tendency of Varieties to depart indefinitely from the original type; and his essay contained exactly the same theory as mine"* (Darwin, 1887).

Both ideas were simultaneously presented, in 1858, in the Journal of the Proceeding of the Linnaean Society (Darwin, 1887), impelling Darwin to finish and publish in 1859 *The Origin of Species by the means of natural selection or the preservation of the favored races in the struggle for life,* a work that lay down the foundations of modern evolutionary biology (Darwin, 1859).

Six years after in 1865, Gregor Mendel's findings from his work with the pea plants would provide the foundations for the emergence of Genetics, introducing the principals of heredity which were read in the meeting of the Natural Science Society in Brno (Cox, 1999). Later, in 1892, August Weismann demonstrated that inheritance only takes place through gametes, putting an end to Lamarck's theory of inheritance of acquired characteristics (Weismann, 1893) and in 1903 Walter Sutton finds that chromosomes are the basis for the Mendelian inheritance (Sutton, 1902; Sutton, 1903). Yet, only in 1944 when Oswald Avery and colleagues continued the research initiated by Frederick Griffith (Griffith, 1928) was the deoxyribonucleic acid (DNA) discovered (Avery *et al.*, 1944). Now the main focus had shifted to the understanding of this enigmatic molecule. In 1953 Watson, Crick and Rosalin Franklin discovered the structure of DNA revealing that it meets the unique requirements for a substance that encodes genetic information (Crick, 1970). Later, in 1961, the code of life is cracked (genetic code) by Marshall W. Nirenberg and collaborators (Nirenberg *et al.*, 1961; Roberts, 1962). Together these discoveries lead to the publication of the central dogma of molecular biology. In 1970, for the first time, the principle of transfer of genetic information is established (Crick, 1970). In this same year Susumu Ohno publishes his book *"Evolution by gene duplication"* (Ohno, 2013) that would significantly impact how future peers would approach research in evolution.

## I.2 Evolution by gene duplication

Gene duplication has long been recognized as an important mechanism in evolution. One of the first observations of the phenotypical outcome from duplication took place in 1936, when Bridges perceived that the duplication of a chromosomal segment in *Drosophila melanogaster* lead to the "Bar-eye" reduction (Bridges, 1936). Nevertheless it was Ohno's observations on duplication in 1970 and the proposal of the of whole genome duplications that constituted a turning point in evolutionary thinking (Ohno *et al.*, 1968; Ohno, 1970). In his book "*Evolution by gene duplication*", he states that natural selection is a very conservative force and that duplication creates the opportunity for a gene or its duplicate to escape from this force, accumulate mutations which may possibly lead to the acquisition of novel functions (Ohno, 2013). Additionally, based on his observations regarding genome sizes in several species he postulated that the observed differences may be due to duplication, and proposed that a tetraploidization event took place with the emergence of the first vertebrate approximately 500 MYA (Ohno *et al.*, 1968; Ohno, 2013). This hypothesis is known today as the 2R hypothesis.

The term "2R hypothesis" for the description of two rounds of whole genome duplication (2R WGD) was only coined years later appearing in several research articles with the most probable first referral dating to 1996 (Sidow, 1996; Hokamp *et al.*, 2003). Initially, the 2R hypothesis was not readily accepted, instead it was challenged and generated a fair amount of controversy in the late 1990s and early 2000s (Skrabanek *et al.*, 1998; Hughes, 1999; Martin, 2001; Hughes *et al.*, 2003). Yet, with the increasing release of genomic data from numerous species (Fig. 1), several key studies were published (Panopoulou *et al.*, 2003; Dehal *et al.*, 2005; Nakatani *et al.*, 2007; Putnam *et al.*, 2008) supporting the 2R hypothesis, that is today largely accepted.

Briefly, the 2R hypothesis suggests that the vertebrate ancestor underwent two separate rounds of whole genome duplication, 2R WGD, approximately 500 MYA (Putnam *et al.*, 2008). Supporting studies show that gene families in vertebrates are generally constituted by multiple members (up to four) that are paralogous originating from duplication, while the corresponding gene (orthologous) family in invertebrates normally presents one member (4:1 ratio), additionally in several cases this observation can be extended to entire genomic segments related by duplication (Paralogons) (Holland *et al.*, 1994; Meyer *et al.*, 1999; Lundin *et al.*, 2003; Panopoulou *et al.*, 2003; Dehal *et al.*, 2005; Putnam *et al.*, 2008). One of the most noticeable examples of the 4:1 ratio is the *Hox* gene family. This family is organized into tight gene clusters encoding DNA-binding proteins, that regulate the segmental structures in embryonic development in bilaterians (Holland, 1992; Amores *et al.*, 1998; Hoegg *et al.*, 2005; Holland, 2013). It was found that invertebrates present generally 1 *Hox* gene cluster while vertebrates species generally tend to present at least 4 *Hox* gene clusters, with the exception of some lineages that underwent additional lineage specific genome duplication (discussed below) (Holland, 1992; Hoegg *et al.*, 2005; Putnam *et al.*, 2008; Holland, 2013).

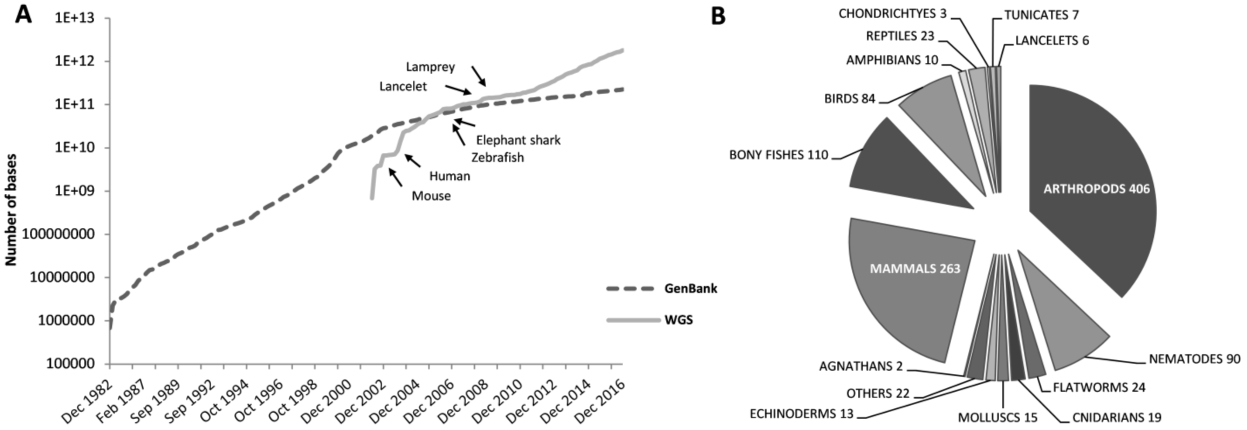


**Figure 1: A**- Time scale of the number of nucleotide bases deposited in NCBI through direct submissions in GenBank and from Whole Genome Shotgun (WGS) sequencing. Arrows indicate approximate time of the first release of WGS for the indicated species. **B**- Distribution of the main metazoan groups with whole genome shotgun sequencing project available in NCBI (data retrieved from NCBI on January 2017).

Despite the general acceptance that 2 rounds of whole genome duplication occurred in the vertebrate ancestor, the exact timing and extension of each round of duplication still remains a matter of debate (Kuraku *et al.*, 2009; Smith *et al.*, 2015) (Fig. 2). This dispute stemmed when several gene families from the vertebrate lineage agnatha were found not to comply with the typical 4:1 ratio (Kuraku *et al.*, 2009). Although initial analysis supported that both rounds of duplication occurred consecutively before the divergence of the agnatha lineage (Kuraku *et al.*, 2009), the release of both sea lamprey (Smith *et al.*, 2013) and Japanese lamprey (Mehta *et al.*, 2013) genomes challenged this assumption.

Still the majority of the evidence supports that the gnathostome lineage diverged after the 2R WGD. However, if both rounds of duplication occurred before the divergence of the agnathan lineage or alternatively if one round occurred before and the second after the divergence of agnathans still remains to be fully resolved (Fig. 2).

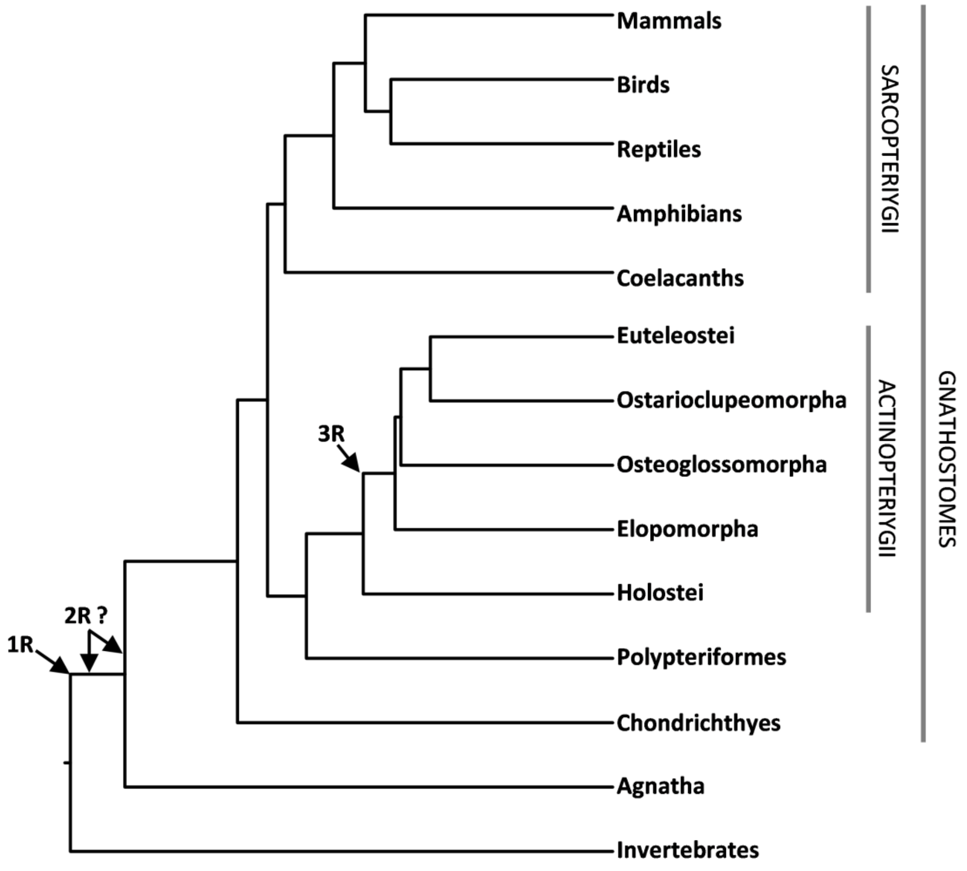


**Figure 2: A**- Phylogenetic tree showing all major chordate lineages, with the approximate indication of timing of whole genome duplications events, 1R first round of whole genome duplication, 2R second round of whole genome duplication, 3R teleost specific genome duplication. Tree calculated at TimeTree public knowledge-base (Hedges *et al.*, 2006; Hedges *et al.*, 2015).

Aside from the whole genome duplications in the vertebrate ancestor, public access to novel genome data increased the identification of additional episodes of genome duplication in vertebrate lineages and species, for example: the teleost specific genome duplication (3R WGD) that occurred approximately 450 MYA (Fig. 2) (Jaillon *et al.*, 2004). This genome duplication took place in the actinopterygii lineage after the divergence of the holostei lineage (Amores *et al.*, 2011) followed by an additional duplication documented in the salmonid lineages (4R WGD) occurring approximately 88 MYA (Moghadam *et al.*, 2011; Macqueen *et al.*, 2014). Additional independent specific duplications were also documented in the ray-finned paddle fish (Crow *et al.*, 2012), in the amphibian *Xenopus laevis* (African clawed frog) (Session *et al.*, 2016) and in *Tympanoctomys barrerae* (red viscacha rat) (Gallardo *et al.*, 1999) the only mammal so far recognized to have undergone a genome duplication event.

## I.3 CONSEQUENCES OF DUPLICATION

Duplication has long been referred as one of the chief driving forces of phenotypic innovation, playing a crucial role in generating variability among species and in evolutionary adaptation (Ohno, 1970; Shimeld *et al.*, 2000; Cañestro, 2012; Chen *et al.*, 2013). In light of Ohno´s view, gene duplication generates a redundant gene that is relieved from selective constraint, allowing it to accumulate mutations without impairing fitness (Ohno, 1970). However, several studies have demonstrated that duplicated genes are not completely freed from selective constraint after duplication; instead duplicate genes may be subjected to alternative selective regimes. For example, it was shown that 17 duplicate genes in *Xenopus laevis* were subjected to purifying selection (Hughes *et al.*, 1993), while other studies have shown that positive Darwinian selection may act after duplication promoting residue variability (Zhang *et al.*, 1998; Lynch *et al.*, 2000; Kondrashov *et al.*, 2002).

Nevertheless, for a duplicate gene to be retained, this gene has to persist in the genome until fixed in the population, resisting loss by mutational inactivation (Innan *et al.*, 2010). In this sense, several evolutionary outcomes have been documented for duplicates genes (Fig. 3) (Force *et al.*, 1999; Lynch *et al.*, 2000; Zhang, 2003; Louis, 2007). After duplication, one duplicate may accumulate several mutations that lead to erosion and ultimately loss – non-functionalization (Fig. 3B) or; one gene copy maintains the ancestral function while the second copy functionally diverges acquiring a novel function – neo-functionalization (Fig. 3C). Alternatively both gene copies may functionally diverge presenting functions that can overlap or complement the ancestral function – sub-functionalization (Fig. 3D); or even, both gene copies maintain the ancestral function but are differentially regulated being expressed in alterative tissues or developmental stages – differential regulation also known as, sub-functionalization through *cis* elements (Fig. 3E) (Lynch *et al.*, 2000; Louis, 2007; MacCarthy *et al.*, 2007). Finally, the preservation of both gene copies expressed simultaneously in the same tissues without functional divergence has also been documented in cases where the gene product is generally in high demand (Fig. 3F). Examples of the later includes ribossomal RNA and histone genes (Zhang, 2003; Scienski *et al.*, 2015). It was proposed that downregulation of the transcription levels of both genes played an important role in the preservation of both redundant copies, assuming that the deletion of either duplicate gene would be disadvantageous, given that only one copy cannot fulfill the required expression level, thus creating a selective pressure for the preservation of both copies (Qian *et al.*, 2010).

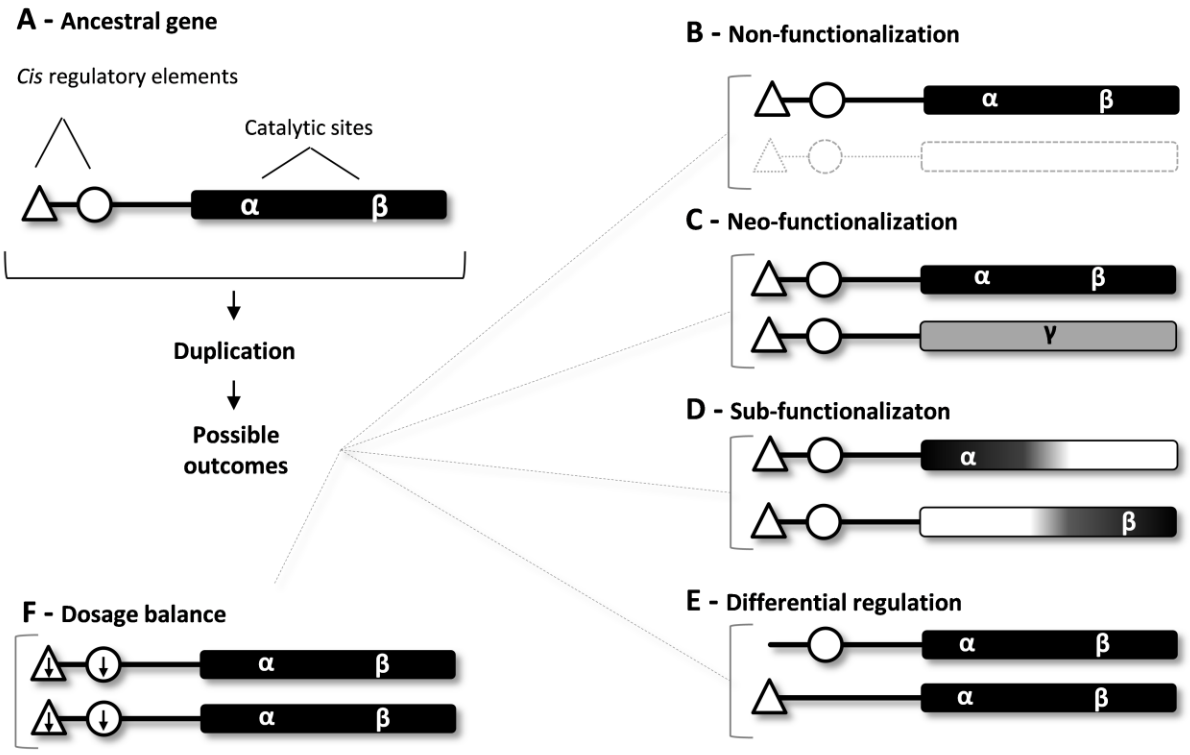


**Figure 3:** Illustration of the several possible outcomes after gene duplication **A**- Schematic representation of an ancestral vertebrate gene containing two active sites α and β and regulated upstream by two *cis* elements Δ and ○ that undergoes duplication. **B**- non-functionalization one duplicate accumulates deleterious mutations that leads to pseudogenization; **C**- neo-functionalization, one copy acquires a novel function and active site γ, **D**- sub-functionalization both copies are maintained by partitioning the ancestral function; **E**- Differential regulation, both copies maintain ancestral function however are expressed in different circumstances and **F** both copies are maintained as well as the ancestral function however, gene expression is downregulated ↓ (adapted from Lynch *et al.*, 2000, Louis, 2007).

Generally, the most frequent outcome for a duplicate gene is degeneration followed by loss due to the accumulation of deleterious mutations in one copy, while the other maintains the original function (Lynch *et al.*, 2000; Kondrashov *et al.*, 2002; Zhang, 2003; Huang *et al.*, 2010). In fact, massive gene loss has been observed occurring shortly after the 2R WGD and teleost specific 3R WGD (Cliften *et al.*, 2006; Huang *et al.*, 2010; Inoue *et al.*, 2015).

## I.4 Evolution by gene loss

Although gene duplication has been viewed as the major source of evolutionary innovation, recent studies have highlighted the crucial role played by gene loss in evolution (Olson, 1999; Albalat *et al.*, 2016). Currently, there are two main evolutionary models for gene loss. The first model proposed by Olson *et al.* - "*Less is more*" - considers gene loss to be an adaptive trait (Olson, 1999). In this sense, the loss of a certain gene or gene family should be advantageous within a particular environmental setting. This type of adaptive evolution can be found in human evolution. The *MYH16* gene suggested to be involved in the head anatomy, is highly expressed in cheek muscles of primates possibly to accommodate a tougher chewing diet (Stedman *et al.*, 2004). Nevertheless it was found to be lost by frameshift mutation in the human lineage, thus removing the anatomical constraints in the head and allowing the development of the modern human brain (Stedman *et al.*, 2004). Advantageous gene loss can also be found in immune response for example: the polymorphic deletion of 32 nucleotides inactivates human CCR5; this gene acts as a receptor for HIV, with homozygous individuals for the deletion being protected against infection by HIV (Dean *et al.*, 1996). The increasing release of genomic data has also allowed the identification of adaptive gene loss in other species. For example, in cetaceans the adaptation to the aquatic environment was followed by the loss of genes involved in hair growth and remodeling of the skin, improving hydrodynamics and reducing drag (Nery *et al.*, 2014; Oh *et al.*, 2015).

The second evolutionary model considers that gene loss events bring no effect to the species involved; here characteristics that have been rendered useless overtime are lost with neutral effects. This model is also known as regressive evolution (Jeffery, 2009; Albalat *et al.*, 2016). A well-known example of regressive evolution derived from environmental conditions can be found in the *Astyanax mexicanus* (cave fish) population. This population is divided into two groups, cave dwelling tetra and surface tetra. Interestingly, it was found that the cave dwelling tetra are blind and display no pigmentation due to the loss of genes related to eye development and pigmentation, unnecessary for life in the darkness (Jeffery, 2009; Albalat *et al.*, 2016). Although this loss is initially viewed as neutral, it can be argued that this loss comes with an energetic benefit to species in a cave ecosystem, where the absence of producers limit nutrient availability, thus rerouting the energy previously allocated for eye development to support growth and other vital functions (Moran *et al.*, 2015). This phenomena has also been observed in other species that colonize dark environments such as naked mole rats (Kim *et al.*, 2011; Emerling *et al.*, 2014) and bats (Zhao *et al.*, 2009).

## I.5 Genome duplications, impact on phylogenetic and synteny analysis

2R and 3R whole genome duplications were followed by extensive chromosome rearrangement, chromosome fission, chromosome fusion and gene loss (Nakatani *et al.*, 2007; Putnam *et al.*, 2008). These events constitute hurdles to be overcome in reconstructing the evolution of a gene family. The typical approach to determine the evolutionary relationships between a set of genes is to perform a phylogeny calculation and analysis of the resulting tree topology. Currently, there is a vast number of different methods available, to determine phylogenetic relationships between sequences (Higgs *et al.*, 2004; Yang *et al.*, 2012). Generally, all phylogenetic methods are based on the same evolutionary notion: that all organisms at one point in time derived from one common ancestor and therefore share similar genes (homologous), the timing of the divergence is largely reflected in the degree of homology erosion observed between sequences (Lemey, 2009).

Typically, the positioning and grouping of sequences from several species in a gene tree reflects the evolutionary history of the corresponding gene. As mentioned above, genome duplications like 2R and 3R were followed by genomic rearrangement. (Nakatani *et al.*, 2007; Putnam *et al.*, 2008) (Fig. 4A). This resulted in the distinct genetic repertoire observed in vertebrate lineages due to differential paralogue retention and lineage specific duplications (Fig. 4B), that together with accelerated sequence divergence may distort the phylogenetic tree topology obscuring the true evolutionary history. The typical topology expected in a phylogenetic tree when all four paralogous that derived from 2R WGD are maintained is observed in the case of the green genes in Fig. 4C. In the case of the dark blue genes, mammals and birds retained paralogue D while teleost retained paralogue C. If no other information is available the inferred phylogenetic tree indicates that one paralogue either C or D was lost in all lineages (Fig. 4D). These uncertainties can be resolved by examining the corresponding gene *loci* and neighboring genes when available (synteny analysis) (Fig. 4B).

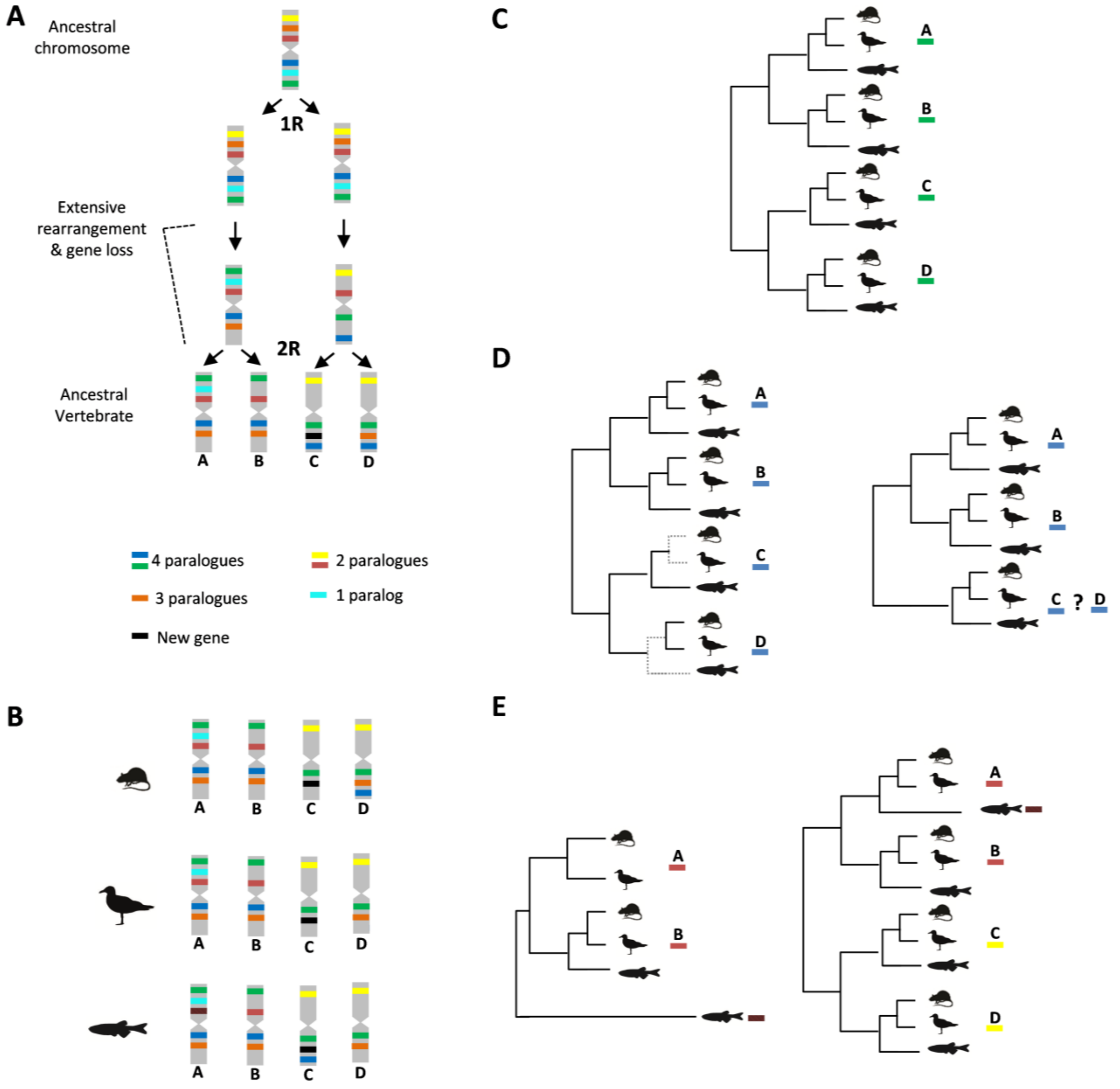


**Figure 4:** Illustration of the phylogenetic and synteny analysis of several genes contained in one ancestral chromosome that underwent 2R WGD **A**- Schematic representation of an ancestral chromosome undergoing 2R WGD, each color bar represents a distinct gene family. **B** – Hypothetical karyotype of 3 animals after 2R WGD. **C** – Phylogenetic analysis of green gene. **D**- Phylogenetic analysis of the dark blue gene family, illustration of differential paralogue retention tree outcomes. **E** – Phylogenetic analysis of the red gene family, illustration of tree outcomes when analyzing divergent sequences with and without outgroup. Adapted from (Kuraku, 2013).

Also, accelerated sequence divergence often camouflages the evolutionary history of a gene, for example the teleost red gene (Fig. 4E). To help clarify these uncertainties and to support phylogenetic analysis a set of sequences from homologous gene family (e.g. yellow genes) should be used as outgroup in the phylogenetic analysis (Fig. 4E).

Additionally, phylogenetic trees are often supported by synteny analysis. Genes related by duplication (paralogous) are placed in genomic *loci* that are also duplicated and in this sense neighboring genes of the target gene are expected to present paralogs in the same genomic *loci* as the target gene. Synteny analysis also indicates if duplicate genes are a result of independent tandem gene duplications and help identify cryptic paralogues or orthologues genes that are poorly placed in the phylogenetic tree due to accelerated sequence divergence or differential paralogue retention.

## I.6 SELECTING RELEVANT SPECIES TO INTERPRET CONTRIBUTIONS OF 2R AND 3R WGD

The selection of species for a project highly depends on the final objective of the study. Studying the impact of genome dynamics on the metabolic pathways in chordates requires a selection of a wide variety of species in order to include all major genomic events and lineages. Besides sampling of all major vertebrate lineages, the work in this thesis has focused on a set of selected species placed in key phylogenetic positions to understand the impact of 2R and 3R WGD.

Cephalochordates (e.g. amphioxus) and tunicates (e.g. sea squirt) diverged from the vertebrate lineage prior to the 2R WGD. In addition, the amphioxus genome presents a considerable amount of synteny conservation when compared to vertebrate genomes consequently, it has been used as a cornerstone in reconstructing the ancestral vertebrate genome (Nakatani *et al.*, 2007; Putnam *et al.*, 2008). Besides the key phylogenetic placing, the amphioxus retains several chordate features such as dorsal nerve cord, notochord, gill slits, segmented muscles, post-anal tail and has been suggested as a model organism for developmental biology (Holland *et al.*, 2004; Holland *et al.*, 2008).

Like the amphioxus, lampreys are also positioned in a key phylogenetic point at the base of vertebrate phylogeny, and have long been viewed as “living fossils” due to conserved morphology observed between living lampreys and fossils found with approximately 360 MYA (Gess *et al.*, 2006; McCauley *et al.*, 2015). Additionally, lipid metabolism in lampreys has been a point of interest in several studies. These studies have focused on lipid accumulation as a decisive factor for the initiation of metamorphosis in juveniles (Lowe *et al.*, 1973; Kao *et al.*, 1997a; Kao *et al.*, 1997b). Together, the extensive knowledge on lamprey biology, life cycle and the availability of whole genome sequencing data from two different species (Mehta *et al.*, 2013; Smith *et al.*, 2013) make the lamprey an attractive model organism to address many aspects of vertebrate evolution (Docker. *et al.*, 2015).

After the divergence of cyclostomes the first gnathostome lineage to diverge was the chondrichthyes, being the first lineage in the vertebrate phylogeny to have fully undergone the 2R WGD. Currently, there are several chondrichthyes genomes available:

*Leucoraja erinacea* (little skate) (Wang *et al.*, 2012), *Callorhinchus milii* (elephant shark) (Venkatesh *et al.*, 2007), *Rhincodon typus* (whale shark) (Read *et al.*, 2015). Chondrichthyes have been used as model organisms to study embryonic development (Cole *et al.*, 2007) and to investigate the evolution of paired appendages in vertebrates (Freitas *et al.*, 2006). Additionally, it was found that chondrichthyes lipid metabolism presented some peculiarities. For example, chondrichthyes present an the low or absent β-oxidation in cardiac and skeletal muscle (Speers-Roesch *et al.*, 2010) and, contrary to mammals that store energy in the adipose tissue, sharks store lipids for energy in the liver (Pethybridge *et al.*, 2014).

Regarding the teleost specific 3R WGD, relevant species to be considered here are those belonging to lineages that diverged before the 3R duplication namely: holostei, polypteriformes, acipenseriformes and lineages that diverged shortly after 3R WGD, the elopomorpha and osteoglossomorpha. Full genome data is available for the holostei *Lepisosteus oculatus* (spotted gar) (Braasch *et al.*, 2016), the elopomorpha *Anguilla anguilla* (European eel) (Henkel *et al.*, 2012b) and for the osteoglossomorpha *Pantodon buchholzi* (African butterfly fish) (Martin *et al.*, 2014). The key phylogenetic placing of spotted gar prior to the 3R teleost genome duplication allowed the identification of many tetrapod orthologous genes that were not found in the zebrafish genome due the extensive rearrangements (Amores *et al.*, 2011; Braasch *et al.*, 2016). Thus, spotted gar genome may be used as a guide to identify genes that were lost after 3R WGD (Amores *et al.*, 2011). Regarding the post 3R WGD species it has been proposed that the elopomorha lineage retained many of the 3R duplicate genes (Henkel *et al.*, 2012a; Henkel *et al.*, 2012b; Chen *et al.*, 2015) contrary to the observed in the recently sequenced osteoglossomorpha African butterfly fish (Martin *et al.*, 2014); thus the analysis of these lineages allows a glimpse into the impact of the 3R WGD untouched by the extensive post duplication rearrangements and gene loss observed in clupeomorpha species.

# CHAPTER II

## LIPIDS AND FATTY ACID METABOLISM

# CHAPTER II – LIPIDS AND FATTY ACID METABOLISM

## II.1 LIPIDS AND FATTY ACID METABOLISM

Lipids constitute a diverse group of biomolecules found in all living organisms, which together with proteins and carbohydrates constitute the three major building blocks of life. Their transversal involvement in numerous biological processes makes their study a fundamental task in the comprehension of biological diversity. Lipids play a significant structural role in biological membranes, inflammatory response, reproduction, sourcing and storing energy and in homoeostasis functioning as signal molecules, cofactors, or endogenous ligands for the nuclear receptor peroxisome proliferator-activated receptor (PPAR) also known as, the master regulator of lipid metabolism (Robinson *et al.*, 2013; Grygiel-Górniak, 2014; Wall *et al.*, 2014).

A distinctive feature of lipids is their insolubility in water. When challenged with a polar environment such as water, lipids will cluster up and expose their polar groups to the environment, shielding the nonpolar carbon-hydrogen chain (Lehninger *et al.*, 2008). This spontaneous assembly of lipids into clusters also known as micelles, constitute the fundamental underpinning for the structure of cellular membranes (Lehninger *et al.*, 2008). Lipids are primarily constituted by fatty acids (FA); these essentially consist of an aliphatic chain ranging from 4 to 36 carbons with a methyl group (Tocher *et al.*, 2015). Fatty acids may be grouped according to the length of the carbon-hydrogen chain into short ($C_{2-}C_4$), medium ($C_6$-$C_{12}$), long ($C_{14}$-$C_{18}$) and very long ($C_{20}$ or more) and by the presence (unsaturated) or absence (saturated) of double bonds in the hydrocarbon chain (Tocher *et al.*, 2015). A small group of lipids do not contain FA. These are essentially cholesterol and other sterols (Lehninger *et al.*, 2008). Aside sterol lipids a great variety of complex molecules are obtained from the elaboration of FA, and these can be grouped into the following major categories: glycerolipids, glycerophospholipids, sphingolipids, prenol lipids, saccharolipids and polyketides (Lehninger *et al.*, 2008).

Fatty acid metabolism encloses several metabolic pathways such as hydrolysis; activation; β-oxidation; biosynthesis; esterification; phospholipid hydrolysis; triglycerides hydrolysis; cholesterol ester synthesis and cholesterol esters degradation.

Generally, research on FA and lipid metabolism has focused in two main fields, human health and aquaculture. A considerable amount of the research in human health is related to obesity, cancer, diabetes and metabolic disorders (Morino *et al.*, 2006; Fucho *et al.*, 2016; Röhrig *et al.*, 2016). Regarding lipid metabolism research in aquaculture, the focus here is shifted to characterizing lipid content, nutritional requirements and metabolism in several cultured species (Tocher, 2010; Tocher *et al.*, 2015). This characterization is of crucial importance given that the aquaculture industry is currently being compelled into more sustainable practices, replacing marine fish meal and oils with plant derived products (Tocher, 2010). Additionally, fish is one of the main dietary sources of highly unsaturated FA omega-3 (ω-3) and omega-6 FA (ω-6) in the human diet, therefore a clear understanding of cultured fish lipid metabolism and requirements is necessary to develop sustainable and cost effective aquaculture (Tocher, 2010; Tocher *et al.*, 2015). Here the integration of *omics* approaches to address nutritional and aquaculture practices has been consistently growing (Castro *et al.*, 2016). For example, comparative genomics is a powerful tool for the identification of genes involved in lipid metabolic pathways, and for the detection of unique genetic repertoires behind alternative metabolic networks observed within the different vertebrate lineages (Zhang *et al.*, 2013; Jiang *et al.*, 2014). Fatty acid metabolism and composition varies among chordates, this variation likely results from the interaction of several factors such as dietary preferences, trophic level, environmental settings and distinct metabolic capabilities (Tocher, 2010; Castro *et al.*, 2016). The ability to endogenously process and elaborate FA is tightly linked with the genetic repertoire of genes involved in FA metabolism, and it has been documented that distinct vertebrate lineages present distinct genetic repertoires (Castro *et al.*, 2012c; Castro *et al.*, 2016). Thus, understanding the evolutionary history of gene families involved in FA metabolism is crucial, as is the link between genetic repertoires and life history trajectories, colonization of new habitats and/or access to new food sources.

Vertebrate evolution is punctuated by events of genome duplication and gene loss (Holland *et al.*, 1994; Nakatani *et al.*, 2007; Holland *et al.*, 2008). Several examples where genome dynamics (e.g. duplication, loss and mutation), environmental factors and diet have modulated lipid metabolism can be found in vertebrates. For example, cats have a limited capacity of endogenously synthesizing arachidonic acid (ARA) due to a limited Δ6 desaturase capacity, a potential consequence of their carnivorous diet that allowed attaining this long chain polyunsaturated fatty acids (LC-PUFA) in sufficient quantities (Rivers *et al.*, 1975; Hassam *et al.*, 1977; Tocher, 2003; Trevizan *et al.*, 2012). Another example of lipid metabolism modulation by dietary preferences can also be found in the human populations. Nordic Inuit populations depend on an extreme diet deprived of fruits, vegetables and grains, and highly rich in ω-3 PUFAS from fatty meat and fish (Fumagalli *et al.*, 2015). It was proposed that this dietary habit resulted in a fixation of specific *alleles* in the FA desaturases, affecting the LC-PUFA biosynthesis pathway in the this population (Fumagalli *et al.*, 2015). On the other hand, the fixation of a distinct allele also known as the *"vegetarian allele"* in the desaturase gene cluster, was observed in African and Asian populations and it was proposed that this *allele* enabled these populations to efficiently convert FAs from plants, medium chain polyunsaturated fatty acids (MC-PUFAS) into LC-PUFAS (Mathias *et al.*, 2012; Kothapalli *et al.*, 2016). Additionally, Inuit populations also present a carnitine palmitoyltransferase 1A gene (*CPT1A*) variant that allows for increased FA oxidation (Collins *et al.*, 2010). This is particularly important given that the Inuit diet is rich in fat content obtained from large artic animals and low in glucose. Therefore, this sequence variant favors energy production via dietary FA oxidation (Wang *et al.*, 2014).

The LC-PUFA biosynthesis pathway has also been modulated in teleost fish by episodes of gene loss. Here, the lack of Δ5 desaturase activity is due to the loss of *fads1*, nevertheless this loss apparently has no significant consequence in marine species given that these species easily obtain LC-PUFA docosahexaenoic acid (DHA) through diet in a DHA rich marine ecosystem (Li *et al.*, 2010b; Tocher, 2010). Alternatively, the loss of *fads1* has been shown to be bypassed in some freshwater or herbivores teleost species who display f*ads*2 desaturases with alternative subtract preferences, and capable of Δ5 desaturations (Hastings *et al.*, 2001; Zheng *et al.*, 2004; Castro *et al.*, 2016).

Besides diet and the genetic repertoire, environmental factors also play a relevant role in the modulation of lipid metabolism. For example, the adaptation to low temperatures has revealed a large number of LC-PUFAs in cell membranes in order to guaranty membrane fluidity (Finegold, 1986). Exposure to low temperatures in humans has also been shown to increase the deposition of brown adipose tissue (Lee *et al.*, 2014).

Therefore, understanding the interplay between genetic repertoire, diet and environment, as well as the reconstruction of the evolutionary history of several gene families involved in FA metabolism such as: FA activation, β-oxidation, FA biosynthesis (Table 1) was the starting point for the development this dissertation.

**Table 1:** Main FA metabolic pathways observed in vertebrate species and corresponding genes involved in each pathway. Underlined gene symbols in the table correspond to gene families investigated during the elaboration of this thesis.

| Process | Gene families |
|---|---|
| FA activation | *Acsl, Acss, Acsbg* |
| FA biosynthesis | *Fasn, Acac, Elovl, Fads, Scd* |
| β-oxidation | *Cpt, Crat Acox, Abcd, Crot* |
| FA hydrolysis | *Acot* |
| FA esterification to TG and PL | *Dgat, Mogat, Gpam* |
| Phospholipid Hydrolysis | *Lipases, Pla1a* |
| Triglycerides Hydrolysis | *Lpl, Lipc, Lipg, Liph* |
| Cholesterol ester synthesis | *Soat, Lcat* |
| Cholesterol ester degradation | *LipA, LipE* |

## II.2 FATTY ACID ACTIVATION

Fatty acid activation is an essential step in FA metabolism precluding many other anabolic and catabolic processes. Generally, FAs are not biologically active and require activation by the fatty acyl-Coenzyme A (Acyl-CoA) before enrolling in processes, such as β-oxidation or esterification into complex lipids (Watkins, 1997). FA activation is catalyzed by acyl-CoA synthetase (ACS) and consists in a two-step thioesterification reaction resulting in a thioester with coenzyme A (CoA) (Watkins *et al.*, 2007). Fatty acid activation was recognized for the first time in 1948 and was referred to as "sparking" or "priming" at the time. Although the molecular process behind was unknown, it was documented that fatty acids required to be "sparked" or activated before enrolling in β-oxidation (Fig. 5) (Grafflin *et al.*, 1948; Knox *et al.*, 1948).

Twenty-six genes coding for ACS enzymes have been documented in the human genome (Watkins *et al.*, 2007). These may be organized into six groups according to the degree of unsaturation and chain length of the FAs favored as substrate: the short-chain ACS-Family (*ACSS*), medium-chain ACS-Family (*ACSM*), long-chain ACS-Family (*ACSL*), very long-chain ACS-Family (*ACSVL*), Bubblegum ACS-Family (*ACSBG*) and ACS-Family (*ACSF*) (Watkins *et al.*, 2007; Soupene *et al.*, 2008). Although some substrate preference overlap is observed, these enzymes differ in tissue distribution and subcellular location, an indication of their highly specific role in FA metabolism (Watkins, 1997).

Short chain FAs play a relevant role in energy homeostasis, appetite regulation, weight, insulin sensing and are the principal fermentation product of non-digestible carbohydrates by the intestinal microbiota (Byrne *et al.*, 2015; Canfora *et al.*, 2015; Morrison *et al.*, 2016). In humans short chain FAs have also been reported to modulate skeletal muscle, liver functions and adipose tissue, through lipolysis and adipogenesis (Canfora *et al.*, 2015). In this sense the ACSS enzymes play a critical role by activating short FAs (Watkins, 1997). Similarly to the ACSS enzymes, ACSL enzymes also play a paramount role in FA metabolism, since FAs with 12 to 20 carbons ($C_{12}$-$C_{20}$) are highly prevalent in the diet and are preferentially converted to acyl-CoA by these enzymes (Watkins *et al.*, 2007; Li *et al.*, 2010a).

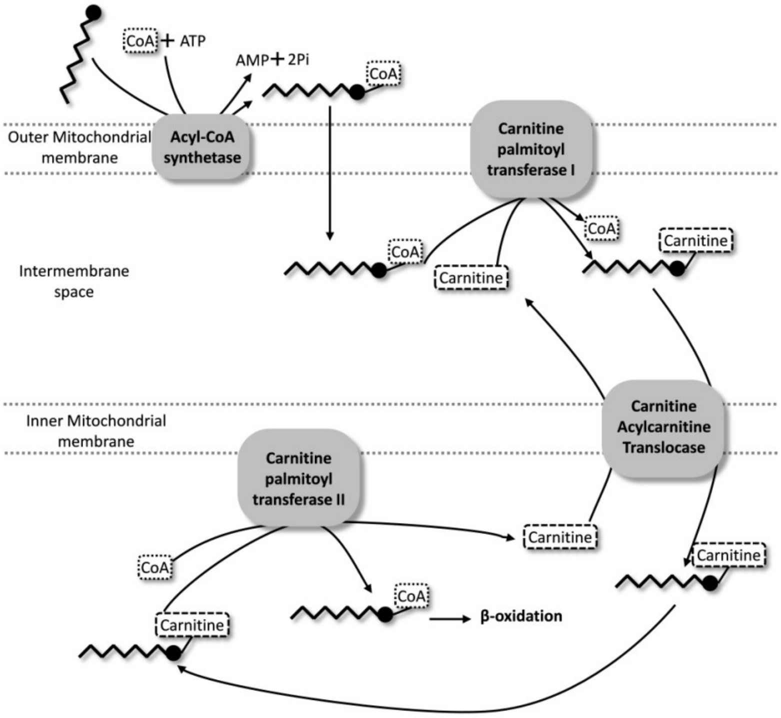

CoA + ATP
AMP + 2Pi
CoA
Outer Mitochondrial membrane
Acyl-CoA synthetase
Carnitine palmitoyl transferase I
CoA
Carnitine
CoA
Carnitine
Intermembrane space
Inner Mitochondrial membrane
Carnitine Acylcarnitine Translocase
Carnitine palmitoyl transferase II
Carnitine
Carnitine
CoA
CoA
β-oxidation
Carnitine

**Figure5:** Schematic representation of FA activation and translocation into the mitochondria for β-oxidation. CoA – Coenzyme A, ATP-Adenosine triphosphate, Pi- inorganic phosphate group. Illustration adapted from (Dunning *et al.*, 2014)

Previous studies identified 3 *Acss* genes *Acss1*, *Acss2* and *Acss3* and five distinct *Acsl* genes in mammals, which were further organized into two separate groups: (i) *Acsl1*, *Acsl5* and *Acsl6*; (ii) *Acsl3* and *Acsl4* (Watkins *et al.*, 2007; Soupene *et al.*, 2008; Li *et al.*, 2010a). Although previous studies have approached the phylogenetic organization and distribution of all ACS enzymes, an explanatory detailed evolutionary history was not provided (Watkins *et al.*, 2007). In this sense, the distribution and evolutionary history of *Acss1* and *Acsl1*, *Acsl3, Acsl4, Acsl5,* and *Acsl6* was revisited and reanalyzed in Chapter III.

### II.3 FATTY ACID BIOSYNTHESIS

Long-chain ($C \geq 20$) polyunsaturated fatty acids such as arachidonic acid (ARA, 20:4n-6), eicosapentaenoic acid (EPA, 20:5n-3) and docosahexaenoic acid (DHA, 22:6n-3) are critical molecules participating in numerous physiological processes such as energy storage, bio-membrane composition and signaling cascades (Tocher, 2003; Schmitz *et al.*, 2008). In addition to the dietary input, LC-PUFAs are endogenously synthesized from essential dietary polyunsaturated FA ($C_{18}$ PUFAS) precursors, including linoleic acid (LOA, 18:2n-6) and α-linolenic acid (ALA, 18:3n-3) (Guillou *et al.*, 2010). This synthesis comprises a series consecutive desaturation and elongation reactions. Typically in mammals, the enzymatic cascade converting $C_{18}$ PUFAs into bioactive LC-PUFAs such as DHA, requires the concerted action of fatty acyl desaturase (FADS) enzymes (FADS1 and FADS2), and the elongase enzymes, for the elongation of very long-chain fatty acids (ELOVL2 and ELOVL5) at specific steps in the pathway (Fig. 6) (Guillou *et al.*, 2010). Therefore, LC-PUFA biosynthesis pathway constitutes an extraordinary example where two unrelated gene families *Elovl* and *Fads* have co-evolved, closely working together for the completion of the pathway. Nevertheless, as referred earlier, this pathway has also been sculpted, by environmental factors and diet, presenting distinct gene repertoires, in several vertebrate lineages (Tocher, 2003; Castro *et al.*, 2016).

The investigation of the LC-PUFA biosynthesis pathway in this thesis was performed in two stages; the first approach was the investigation of the distribution of the *Elovl* gene family in chordates and functional characterization of Elovl enzymes from species placed in key phylogenetic positions. The second approach consisted in a similar investigation of the *Fads* gene family.

The Elovl2 and Elovl5 are fatty acid elongase enzymes found in several vertebrate species, which elongate polyunsaturated fatty acids (PUFAS) by the addition of 2 Carbon molecules at the carboxyl end (Leonard *et al.*, 2004). In humans the Elovl5 presents a substrate preference for dietary fatty acids $C_{18}$ to $C_{20}$, while the Elovl2 presents a substrate preference for $C_{20}$, $C_{22}$ and $C_{24}$ (Leonard *et al.*, 2002; Leonard *et al.*, 2004).

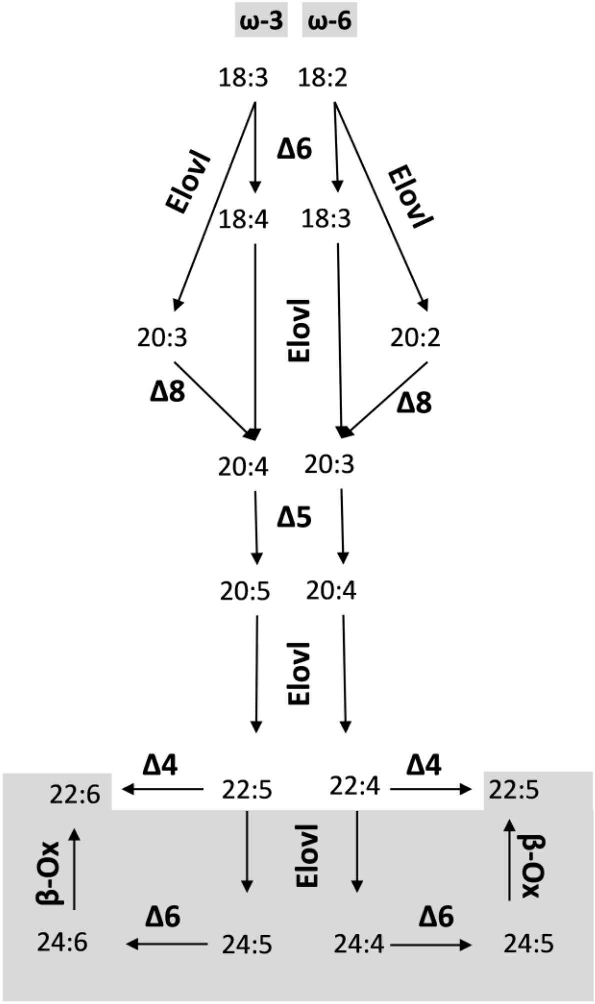


**Figure 6:** Schematic representation of the LC-PUFA biosynthesis pathway, elongation (Elovl), desaturation (Δ4, Δ5, Δ6, Δ8), β-Ox indicates β-oxidation pathway, omega-6 (ω6) and omega-3 (ω3) pathways are depicted in parallel and each fatty acid is represented by a composite number, for example 18:2 corresponds to linoleic acid and grey box indicates the Sprecher pathway (Sprecher, 2000).

Functional overlap and sequence homology between *Elovl2* and *Elovl5* hint towards a common evolutionary origin. While the majority of vertebrate lineages present both elongase genes, the *elovl2* seems to have been lost in the majority of marine and commercial teleost species, impairing the endogenous synthesis of DHA, *via* Spreecher pathway (Morais *et al.*, 2009; Castro *et al.*, 2016).

Given that the completion of the LC-PUFA biosynthesis pathway requires the action of both elongases and desaturases, the next step was to investigate the *fads* gene family.

While Elovl enzymes introduce 2 carbon atoms, Fads enzymes remove hydrogens creating double bonds, were delta (Δ) indicates the position in which the double bond is created, for example Δ6 - 6th position from the carboxyl group (Los *et al.*, 1998). In the LC-PUFA biosynthesis, we find two key desaturase genes, *Fads1* and *Fads2*. These genes are located in a gene cluster were it is also possible to identify an additional *Fads3* gene with no known function reported yet (Marquardt *et al.*, 2000; Blanchard *et al.*, 2011). This disposition indicates that the *Fads* genes arose from independent tandem duplication events. However, the timing and distribution of *Fads* genes in vertebrates remains to be resolved. Initial investigations suggested that *Fads* diversification took place before the divergence of the mammalian lineage. However the identification of *Fads1* and *Fad2* orthologues in chondrichthyes indicated an older origin of the desaturase genes, revealing an alternative evolutionary history with the loss of *Fads1* in the teleost lineage (Castro *et al.*, 2012c). Similarly to the loss of *Elovl2*, the loss of *Fads1* in teleosts also impairs the LC-PUFA biosynthesis. However, in some species such as *Danio rerio,* Fads2 presents alternative activities such as Δ5/Δ6/Δ8 compensating the loss of Fads 1 (Hastings *et al.*, 2001; Monroig *et al.*, 2011a). Yet, the exact timing of *Fads1* loss in the teleost lineage remains unclear, as does the desaturase gene complement in the basal vertebrate lineage, the cyclostomes. The evolutionary history, distribution of both elongases and desaturases involved LC-PUFA biosynthesis is examined in Chapter IV.

## II.4 β-OXIDATION

After FA activation and/or FA biosynthesis, FAs may undergo catabolism through the β-oxidation pathway in the mitochondria, thus playing a significant role in energy homeostasis. Each cycle of β-oxidation shortens the FA by 2 carbons, producing acetyl–CoA, which is the primary substrate used in the Krebs cycle for energy production (Lehninger *et al.*, 2008). The import of long chain FA into the mitochondria is mandatory for β-oxidation. This process is mediated by the carnitine acyltransferase (CPT) system composed of CPT1 and CPT2 (McGarry *et al.*, 1997), that catalyze the reversible exchange in FA of Coenzyme-A (CoA) and carnitine. This is a two-step process. Carnitine is bound to FA by the action of CPT1 with the release of CoA. Then the FA bound to carnitine transverses to the inner mitochondrial membrane. Next, CPT2 reverses this exchange by releasing the carnitine and reattaching a CoA group to the FA, which then may undergo to β-oxidation (Fig. 5). Since the inner membrane is only permeable if the FA is linked to carnitine, CPT1 assumes a central rate limiting role in this pathway (McGarry *et al.*, 1997). In mammals, the *Cpt1* gene family is encoded by three separate genes designated *Cpt1a*, *Cpt1b* and *Cpt1c* (Esser *et al.*, 1996; Van der Leij *et al.*, 2000; Bonnefont *et al.*, 2004; Boukouvala *et al.*, 2010). Importantly, the evolutionary history and distribution of the *Cpt1* genes has posed complex questions. That is specially the case of *Cpt1c*, which has been considered a recent duplicate originated in the ancestor of mammals (Boukouvala *et al.*, 2010; Lee *et al.*, 2012). In Chapter V the evolutionary history of this gene family is analyzed.

Saturated FA with an even number of carbons atoms are degraded *via* the β-oxidation pathway; however the completion of this pathway with unsaturated FA or/and FA with an odd number of carbons atoms requires additional steps (Lehninger *et al.*, 2008). The oxidation of an FA containing an odd number of carbons atoms yields a propionyl-CoA and a acetyl-coA rather than two acetyl-CoA (Lehninger *et al.*, 2008). In order to metabolize propionyl-CoA in the Krebs cycle for energy production, propionyl-CoA has to be converted into succinyl-coA. This molecular rearrangement is performed by methyl-malonyl-CoA mutase, which requires vitamin B12 as a coenzyme (Smith *et al.*, 1999). Vitamin B12 deficiency leads to the toxic accumulation of propionyl-coA, severe neurological dysfunction and anemia (Briani *et al.*, 2013). Animals are unable to endogenously synthesize B12 and therefore rely on dietary B12. However, the safe passage of B12 in the route from ingestion, to intestinal absorption and finally to the conversion of propionyl-CoA, depends on a complex relay of cobalamin binders (Fedosov *et al.*, 2007; Greibe *et al.*, 2012). In humans, three cobalamin binders encoded by 3 genes, *Gif* (gastric intrinsic factor), *Tcn1* (haptocorrin), and *Tcn2* (transcobalamin), have been identified (Fedosov *et al.*, 2007; Quadros, 2010). This diversity of binders assures that B12 is efficiently transported through several anatomical and physiological environments. The observed distribution of the cobalamin binder genes within vertebrates lead to the proposition that the diversification of cobalamin binders only occurred in the tetrapod lineage, given that teleost species presented one cobalamin binder resembling the human binders (Greibe *et al.*, 2012). Further investigation of this gene family described in Chapter V revealed an alternative evolutionary history of this gene family.

### II.5 GASTRIC PROTEASES AND PROTEIN DIGESTION

Proteins are the most abundant biomolecule found in living organisms, occurring in all cells, participating in numerous functions and mediating countless biological processes. Proteins are involved in structural functions (e.g. keratins, muscle fibers), they also play a relevant role in immune response (e.g. antibodies), transport (e.g. hemoglobin), sensing (e.g. photoreceptor proteins), and in global metabolic roles with enzymes and hormones (Lehninger *et al.*, 2008).

Protein digestion constitutes yet another remarkable example of the interplay between gene repertoire, diet and metabolism. For example, the domestication of dogs entailed the reshaping of its digestive enzymes to efficiently digest starch. While wolves (exclusive carnivores) present two amylase genes, the domestic dog displays a considerably higher number of copies which varies from 4 to 30 genes (Axelsson *et al.*, 2013). Also, it has been proposed that the expansion of the amylase 1 gene (*AMY1*) in humans occurred in populations that present a high starch diet (Perry *et al.*, 2007). Similarly, gastric enzymes such as Pepsinogen C (*PgC*), Pepsinogen B (*PgB*), Pepsinogen A (*PgA*), and Chymosin (*Cmy*) have also been shown to vary across vertebrate lineages. For instance, *PgA* presents linage specific duplication and loss within hominids (Ordoñez *et al.*, 2008; Narita *et al.*, 2010). It has also been demonstrated that the genetic repertoire of the gastric genes and proteolytic activity can be correlated with diet (Chan *et al.*, 2004), and indicate the presence or absence of a stomach in gnathostomes (Castro *et al.*, 2014). Rerouting of pepsinogens towards other functions has also been observed in the pufferfish, which does not have a stomach while retaining a pepsinogen expressed in the skin, suggesting that this enzyme in this species has acquired an alternative function to food digestion (Kurokawa *et al.*, 2005).

Within all gastric enzymes, we find that the pepsin gene family generally consists of 5 members grouped according to sequence identity and substrate specificity: *Cmy, PgA, PgB, PgC* and Pepsinogen F (*PgF*), which are highly expressed in the gastric mucosa (Yakabe *et al.*, 1991; Kageyama, 2002; Carginale *et al.*, 2004; Wu *et al.*, 2009). These proteases are secreted generally as inactivate zymogens that then undergo autocatalytic activation in the gastric tract (Richter *et al.*, 1998). Although this gene family is widely disseminated, it is unevenly distributed within gnathostomes, presenting cases of gene expansion (e.g. *PgA* in primates and lagomorpha), pseudogenization and loss (Kageyama, 2002; Castro *et al.*, 2014). Independent events of gene expansion in specific vertebrate lineages have been suggested to allow the emergence of gastric proteases presenting different substrate specificities (Narita *et al.*, 2010). It was previously suggested that the different types of pepsinogens (*PgA, PgF, PgC, PgB* and *Cmy*) evolved from a common ancestor aspartic protease (Kageyama, 2002). Additionally, *PgC* was also suggested to be a suitable molecular marker in vertebrate phylogeny given its single copy status in vertebrates (Kageyama, 2002). Nevertheless, the evolutionary history of pepsinogen family was not yet consolidated. In this thesis, the evolutionary history of two pepsin gene families *PgC* and *Cmy* was examined in Chapter VI.

# Objectives

# OBJECTIVES

Several studies have revealed the impact of gene/genome duplication, gene loss and mutation in numerous vertebrate gene families, enlightening the key role played by these processes in the evolution of vertebrate diversity, physiology and adaptation (Holland *et al.*, 1994; Shimeld *et al.*, 2000; Glasauer *et al.*, 2014).

The evolutionary history of gene families involved in FA metabolism and protein digestion in vertebrate species has been previously investigated, uncovering variable genetic repertoires, alternative metabolic capabilities and pathways, supporting the key role of 2R WGD, 3R WGD, gene duplication and loss in the sculpting of these metabolic pathways (Castro *et al.*, 2012c; Castro *et al.*, 2014; Castro *et al.*, 2016). Nevertheless, many gene families involved in these pathways are yet to be thoroughly examined and a unifying cross species view of the impact of these genomic processes remains to be portrayed.

Therefore, the global aim of this thesis is to contribute for the understanding of how these evolutionary processes have sculpted the elaboration and diversification, of FA metabolism and protein digestion in vertebrate history. Furthermore, this work aims to comprehend if and how distinct gene repertoires observed in vertebrates are linked to diverse FA profiles and FA physiological metabolizing capacities. Also, this thesis aims to incorporate the evolutionary histories of gene families, with information regarding vertebrate phenotypic diversification, adaptation to novel environmental settings and dietary habits in an effort to contextualize my findings.

To this end, the main focus is to investigate what was the role played by duplication events such 2R (invertebrate/vertebrate transition) and 3R (pre-dating teleost radiation) in the elaboration of the following pathways: FA activation, FA biosynthesis, β-Oxidation and protein digestion, and to clarify the evolutionary history for each gene family studied in several vertebrate lineages. Here, taking advantage of the increasing genomic data accessible in public databases and, several key species available in CIIMAR, targeted gene families will be searched retrieved and identified. Next, sequences sampled or isolated will be characterized to identify key features, used for phylogenetic calculation, and when possible functionally characterized.

To achieve these main goals, a series of specific objectives was set:

1- Identification and characterization of the genetic repertoire involved in short chain FA activation in vertebrates namely *Acss1*;
2- Reexamination of the genetic repertoire involved in long chain FA activation in vertebrates, clarification of the genomic events supporting the grouping of ACSL enzymes into two distinct groups (i) *Acsl1 Acsl5 Acsl6* and (ii) *Acsl3 Acsl4*;
3- Analysis of the contribution of 2R WGD or/and 3R WGD for teleost specific FA activation enzymes;
4- Analysis of the genetic repertoire of Elovl and Fads enzymes involved in LC-PUFA biosynthesis in vertebrates;
5- Delineate the evolutionary time frame in which the complete LC-PUFA pathway emerged through the isolation and functional characterization of Elovl and Fads enzymes from various vertebrate species;
6- Analysis of the impact of 2R WGD on the diversification of the genetic repertoire of *CPT1* gene family in vertebrates;
7- Reexamination of the evolutionary history of B12 binders in vertebrates;
8- Investigation and characterization on the genetic processes involved in the expansion of the *pepsinogen C* gene family in vertebrates;
9- Examination of the evolutionary history in mammals of neonatal protease Chymosin established pseudogene in human.

# CHAPTER III

## FATTY ACID ACTIVATION

# Chapter III – Fatty Acid Activation

## III.1 A novel acetyl-coA synthetase short-chain subfamily member 1 (*Acss1*) gene indicates a dynamic history of paralogue retention and loss in vertebrates

L. Filipe C. Castro*, Mónica Lopes-Marques*, Jonathan M. Wilson, Eduardo Rocha,

Maria A. Reis-Henriques, Miguel M. Santos, Isabel Cunha

(* Joint First authors)

**Gene**

(doi:10.1016/j.gene.2012.01.013)

# A novel Acetyl-CoA synthetase short-chain subfamily member 1 (*Acss1*) gene indicates a dynamic history of paralogue retention and loss in vertebrates

L. Filipe C. Castro [a,*,1], Monica Lopes-Marques [a,1], Jonathan M. Wilson [a], Eduardo Rocha [a,b],
Maria A. Reis-Henriques [a], Miguel M. Santos [a,c], Isabel Cunha [a]

[a] *Interdisciplinary Centre for Marine and Environmental Research (CIIMAR), CIMAR Associate Laboratory, University of Porto (U.Porto), Portugal*
[b] *Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto (U.Porto), Portugal*
[c] *Faculty of Sciences, University of Porto (U.Porto), Portugal*

ARTICLE INFO

*Article history:*
Accepted 14 January 2012
Available online 31 January 2012

*Keywords:*
ACSS1
Paralogue retention
Gene loss
Genome duplications

ABSTRACT

Acetyl-CoA short chain synthetases (ACSSs) are key enzymes in the activation of fatty acids through the formation of thioesters with CoA. Three subfamily members are currently recognized in the human genome, *ACSS1*, *ACSS2* and *ACSS3*, all single copy genes. The mitochondrial isoform, *Acss1*, plays a key role in the metabolism of acetate for energy production. While the single copy condition has been accurately established in humans, the evolutionary history of the *Acss1* subfamily in vertebrates has yet to be elucidated, in particular, the isoform diversity, origin and function. Through genome database mining we analyzed the diversity of *Acss1* isoforms in vertebrate classes. We detected the presence of a novel *Acss1* isoform, which we name *Acss1B*. This new gene, *Acss1B*, has a curious phylogenetic distribution being found in teleosts (except zebrafish), sauropsids (birds and reptiles) and probably chondrichthyes. In contrast *Acss1A* is found in all the investigated species, except the teleost medaka. By means of comparative genomics and phylogenetics we show that *Acss1A* and *Acss1B* were generated in the quadruplication of the vertebrate genome. In effect, we find that amphioxus, a pre-genome duplication chordate, has a single *Acss1* gene in a genomic region equally related to a quadrupled vertebrate genomic set. Consequently, *Acss1B* has been lost in some teleosts, amphibians and mammals, while *Acss1A* is probably absent in medaka. The reported findings illustrate an especially dynamic pattern of paralogue retention and independent loss in vertebrate species involving the *Acss1* subfamily, whose functional consequences in energy metabolism are as yet unknown.

## 1. Introduction

Twenty-six distinct Acyl-coenzyme A synthetase (ACS) proteins are currently recognized in the human genome (Watkins et al., 2007). According to the chain length of their fatty acid substrate ACSs are divided into six families: short-chain (ACSS), medium-chain (ACSM), long-chain (ACSL), very long-chain (ACSVL), bubblegum (ACSBG) and a group of uncharacterized ACSs (ACSF) (Watkins et al., 2007). The existence of such a large portfolio of ACSs suggests that each plays a unique role, directing the acyl-CoA product to a specific metabolic fate (Watkins et al., 2007). Being involved in the activation (formation of a thioester bond with coenzyme A) of short chain fatty acids to form fatty acid acyl-CoA, ACSSs play a particularly vital metabolic function. Their importance has been recognized for decades, since they provide the cells with the two-carbon metabolite used in many anabolic and energy generation processes (Starai and Escalante-Semerena, 2004). In eukaryotes ACSS plays an exclusive role in the activation of acetate and is therefore of critical importance, comparatively to prokaryotes that may present alternative pathways (Starai and Escalante-Semerena, 2004). ACSSs are most active towards acetate, propionate, or butyrate (though the activity for butyrate and propionate is *ca* 25- to 35-fold greater than for acetate) (Orchard and Anderson, 1996).

In humans the ACSS gene family comprises three subfamilies (Watkins et al., 2007). ACSS1 is a mitochondrial enzyme that catalyzes the oxidation of acetate under ketogenic conditions (Fujino et al., 2001; Sakakibara et al., 2009), while the cytosolic isoform ACSS2 is thought to be involved in production of acetyl-CoA from free acetate for the synthesis of fatty acids and sterols (Ikeda et al., 2001). The recently described ACSS3 is found also in the mitochondrion and has not yet been functionally characterized (Pérez-Chacón et al., 2009).

Here we investigate the evolution of the *Acss1*subfamily in vertebrate classes. While its single copy condition has been confidently

*Abbreviation:* ACSSs, Acetyl-CoA short chain synthetases; Acss1, Acetyl-CoA synthetase short-chain family member 1; ACS, Acyl-coenzyme A synthetase; ACSM, Acyl-coenzyme A synthetase medium-chain; ACSL, Acyl-coenzyme A synthetase long-chain; ACSVL, Acyl-coenzyme A synthetase very long-chain; ACSBG, Acyl-coenzyme A synthetase bubblegum; ACSF, Uncharacterized Acyl-coenzyme A synthetase; ORF, open reading frame.

[*] Corresponding author at: CIIMAR, Rua dos Bragas, 289, 4050-123 Porto, Portugal. Tel.: +351 223401800; fax: +351 223390608.
*E-mail address:* filipe.castro@ciimar.up.pt (L.F.C. Castro).

[1] joint first authors.

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.01.013

established in humans and mice (Watkins et al., 2007), no information in other species has yet been gathered. However, the advent of numerous full genome projects covering an ample phylogenetic range has significantly impacted our ability to understand gene isoform diversity. We find a previously undocumented *Acss1* gene, which we name *Acss1B*, present in some teleost species, sauropsids and probably chondrichthyes. Phylogenetics, synteny and paralogy analysis indicate that *Acss1A* and *Acss1B* resulted from genome duplications in vertebrate ancestry. We conclude that the *Acss1* subfamily underwent events of paralogue independent loss and retention following 1R/2R genome duplications.

## 2. Materials and methods

### 2.1. Database searches and identification of Acss1 sequences

The identification of *Acss1* sequences was achieved through Blast searches to the Ensembl and GenBank databases. The full coding sequence of human ACSS1 was used as query for Blastp. Searches were performed in the following species: *Homo sapiens* (human), *Mus musculus* (mouse), *Monodelphis domestica* (grey short-tailed opossum), *Gallus gallus* (chicken), *Anolis carolinensis* (green anole), *Xenopus tropicalis* (western clawed forg), *Danio rerio* (zebrafish), *Oryzias latipes* (medaka), *Gasterosteus aculeatus* (stickleback), *Tetraodon nigroviridis* (green spotted pufferfish), *Ciona savignyi* (sea squirt), *Branchiostoma floridae* (amphioxus) and *Strongylocentrotus purpuratus* (sea urchin). The *Acss1* gene search in *Takifugu rubripes* (pufferfish) was performed in the fifth genome assembly at http://www.fugu-sg.org/BLAST/Blast2.htm. The search to the *Callorhinchus milii* (elephant shark) was performed using the human ACSS1 sequence with tblastn at http://esharkgenome.imcb.a-star.edu.sg/Blast/. Intron/exon predictions were made with FGENESH+(http://linux1.softberry.com/berry.138 phtml).

### 2.2. Phylogenetic analysis

The ACSS1 amino acid sequences retrieved from the genome search were aligned using MAFFT with the L-INS-i method (Katoh and Toh, 2008). Accession numbers of each sequence are indicated in Fig. 1. Gaps were removed to obtain a final alignment of 527 amino acids of 22 sequences. To determine the best model of amino acid substitution we run ProtTest (LG+I+G) (Abascal et al., 2005). Finally, a Maximum Likelihood tree was reconstructed using PhyML online (Guindon et al., 2010), with the proportion of invariant sites calculated from the alignment, and four rate categories with a gamma distribution parameter estimated from the data. Confidence in each node was assessed by 1000 bootstrap replicates of the data. Trees were visualized with TreeView 1.6.6. The tree was rooted with the ACSS1 orthologue from *S. purpuratus*.

### 2.3. Comparative genomics and neighbouring gene families

We collected the information on the three closest gene ORFs flanking *Acss1* genes (or *locus*) using the following genome releases from Ensembl: *H. sapiens*- GRCh37, *G. gallus*-WASHUC2, *A. carolinensis*-AnoCar2.0, *X. tropicalis*-JGI_4.2, *D. rerio*-Zv9, *O. latipes*-MEDAKA1, *G.*

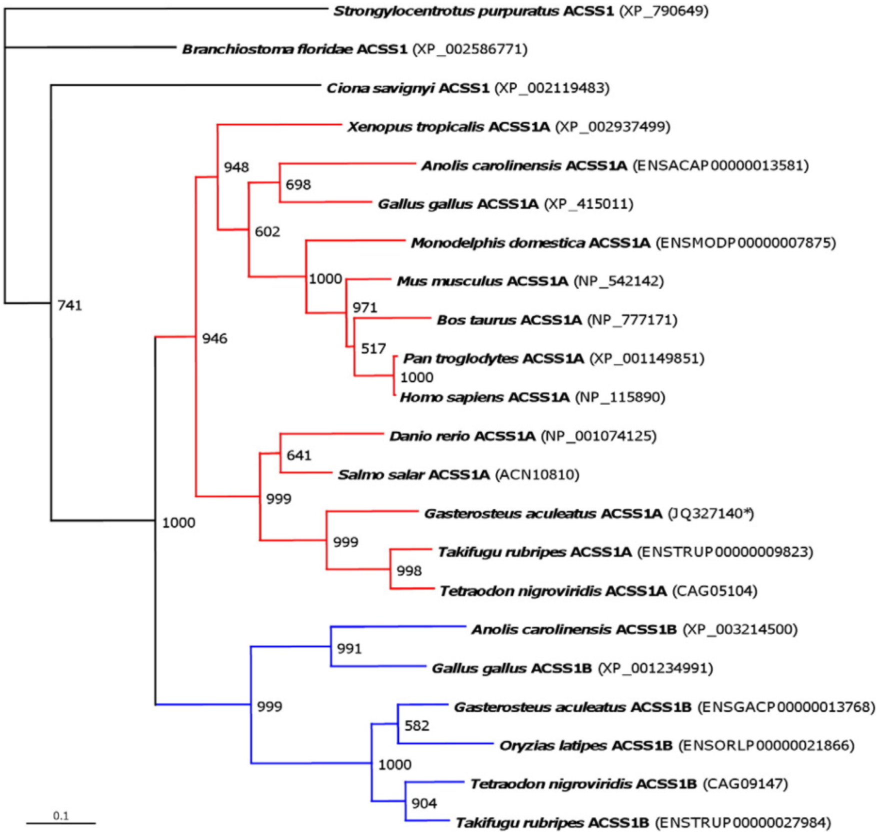

Strongylocentrotus purpuratus ACSS1 (XP_790649)
Branchiostoma floridae ACSS1 (XP_002586771)
Ciona savignyi ACSS1 (XP_002119483)
Xenopus tropicalis ACSS1A (XP_002937499)
948
Anolis carolinensis ACSS1A (ENSACAP00000013581)
698
Gallus gallus ACSS1A (XP_415011)
602
Monodelphis domestica ACSS1A (ENSMODP00000007875)
1000
Mus musculus ACSS1A (NP_542142)
741
971
Bos taurus ACSS1A (NP_777171)
946
517
Pan troglodytes ACSS1A (XP_001149851)
1000
Homo sapiens ACSS1A (NP_115890)
Danio rerio ACSS1A (NP_001074125)
641
Salmo salar ACSS1A (ACN10810)
999
Gasterosteus aculeatus ACSS1A (JQ327140*)
1000
999
Takifugu rubripes ACSS1A (ENSTRUP00000009823)
998
Tetraodon nigroviridis ACSS1A (CAG05104)
Anolis carolinensis ACSS1B (XP_003214500)
991
Gallus gallus ACSS1B (XP_001234991)
999
Gasterosteus aculeatus ACSS1B (ENSGACP00000013768)
582
Oryzias latipes ACSS1B (ENSORLP00000021866)
1000
Tetraodon nigroviridis ACSS1B (CAG09147)
904
0.1
Takifugu rubripes ACSS1B (ENSTRUP00000027984)

**Fig. 1.** Molecular phylogenetic analysis by the Maximum Likelihood method of *ACSS1* amino acid sequences. Sea urchin was used as outgroup. Scale bar denotes 0.1 underlying amino acid substitutions per site. Numbers at nodes indicate robustness estimated from 1000 bootstrap replicates. Accession sequence numbers or ensemble codes are shown for each sequence. *Accession number for the cDNA sequence of GacACSS1A (full assembled sequence in Supplementary Fig. 1B).

*aculeatus*-BROADS1, *T. nigroviridis*-TETRAODON8, and *C. savignyi*-CSAV2.0. Information on the mapping location of *Acss1* in *B. floridae* was collected from GenBank (NZ_ABEP00000000). For *S. purpuratus* we used the assembly Spur_v2.1. In the paralogy analysis we also collected the mapping information of the gene families which are multi-copy using the prediction paralogue/orthologue tool from the Ensembl pipeline.

### *2.4. Identification and characterization of the Acss1A gene in stickleback*

Blast search in stickleback genome retrieved the following hits: one corresponded to a full *Acss1* open reading frame (ORF) and two partial hits located at the limit of two distinct contigs (scaffold 156—contig 13116 and group VI—contig 12400). Both of these hits presented neighbouring genes with a conserved synteny when compared to the localization of *Acss1A* gene in other species (see results section). One corresponds to an annotated partial *Acss1*-like gene (ENSGACG00000001372), the second hit matched to an unannotated region in group VI. Genomic sequence of groupVI-contig12400 was retrieved from Ensemble database for gene prediction in FGENESH + (http://linux1.softberry.com/berry.phtml), using the green pufferfish ACSS1A amino acid sequence as reference. Intron and exon boundaries of the predicted gene were manually curated using the green pufferfish *Acss1A* gene structure as guideline. In order to demonstrate genomic linkage between scaffold 156-contig 13116 and group VI-contig 12400, a set of primers was created using Primer3 software (Rozen and Skaletsky, 2000). One primer was located in an exon of the partial annotated sequence located in scaffold156 (R-5′ GCGTATCTGGCGATCTTTGT3′) and the second in a predicted exon located in group VI (F-5′ GGCTCTTCGTCTCCTGCTAA3′). Using 2 μl of stickleback brain cDNA a PCR was performed with Phusion® Flash high-fidelity Master Mix (FINNZYMES) with the following parameters, initial denaturation at 98 °C for 10 s, followed by 45 cycles of denaturation at 98 °C for 1 s, annealing at 53 °C for 5 s and elongation at 72 °C for 10 s. PCR products were loaded onto 2% agarose gel in TBE buffer and run at 80 V. Gels were stained with GelRed and images acquired. The amplification product was then excised and purified with Illustra GFX PCR DNA and Gel Band purification Kit (GE Healthcare UK) according to manufacturer's guidelines and sequenced directly with the PCR primers (StabVida, Portugal). The cDNA sequence was then assembled with both genomic contigs to demonstrate sequence continuity (Supplementary Fig. 1 A and B).

### *2.5. In silico transcription analysis and RT-PCR*

*In silico* transcription analysis, for *ACSS1A* in human and chicken, was performed using ESTs available from Unigene (Wheeler et al., 2007), as count per million transcripts, all values are displayed as Log2 transcripts per million. A heat map was created using the collected EST data and matrix2png web interface v1.2.1 (Pavlidis and Noble, 2003). Non-quantitative transcription data for *Acss1B* in chicken was retrieved from the BBSRC chickEST database using tblastn (Boardman et al., 2002).

Adult wild-type Zebrafish were obtained from our own breeding stock. Stickleback specimens were caught in Rio Minho, Portugal. Animals were anesthetized and euthanized by cervical transection in accordance with the Portuguese Animals and Welfare Law (Decreto-Lei no 197/96) approved by the Portuguese Parliament in 1996. Institutional animal approval by CIIMAR/UP and DGV (Ministry of Agriculture) was granted for this study. After collection, the tissues (ovary, testis, heart, liver, spleen, kidney, eye, brain, gill, gut, intestine) were preserved in RNAlater and kept at −20 °C. Total RNA was isolated using an Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, UK). All steps were performed according to the manufacturer's instructions, including the on-column treatment of isolated RNA with RNase-free DNase I. RNA concentration was calculated using Qubit fluorometer instrument (Invitrogen, Carlsbad CA), integrity confirmed by electrophoresis and the RNA stored at −80 °C until further use. The cDNA was synthesized from 250 ng of total RNA with the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufacturer's protocol. All samples were further diluted with an equal volume of ultra-pure water. Primers for RT-PCR were designed from gene sequences available in Ensembl with the Primer3 software (Rozen and Skaletsky, 2000). Regarding zebrafish, forward and reverse primers were designed to flank an intron in order to prevent quantification of genomic DNA (F-5′GGACAGATAATAACAGGAAAT 3′; R 5′ TCTGGTAGTATCCGTCTTC 3′; annealing at 52–53 °C). For stickleback, two sets of forward and reverse primers were also designed to flank an intron and prevent quantification of genomic DNA. The first set of primers match the annotated *Acss1B* gene (ENSGACG00000010367) (F-5′ACTTCTTCACTGGGGATGG3′; R-5′CACTGAGGAGGATTGATTC3′; annealing at 58 °C), the second set corresponds to the assembled *Acss1A* (F-5′ CGCCGTTCCTGAGCACTTCCTG3′; R-5′CACCGAGGTCACCTGTGGTCTCC3′; annealing at 55 °C).

PCR was performed with 1 μl of cDNA and the GoTaq system (Promega). The general PCR profile was as follows: an initial denaturation of 2 min at 95 °C, followed by 30 s at 95 °C, 30 s at annealing temperature of the respective primer, 15 s at 72 °C for 35 cycles. Reactions were carried out in a Biometra Personal Thermal Cycler. All PCR products were loaded onto 2% agarose gels in TBE buffer and run at 80 V. Gels were stained with GelRed and images acquired.

## 3. Results and discussion

### *3.1. A novel Acss1 isoform in vertebrate classes*

We used the human ACSS1 (Q9NUB1) sequence for Blastp in various available genomes. In total we investigated 11 species covering a wide range of vertebrate lineages such as eutherian mammals (human and mouse), marsupials (grey short-tailed opossum), birds and reptiles (chicken and green anole), amphibians (western clawed frog), teleosts (green spotted pufferfish, pufferfish, zebrafish, stickleback and medaka). To refine our evolutionary analysis, we also inspected the invertebrate deuterostomes sea squirt, amphioxus and sea urchin.

The search of *Acss1* genes in genome databases found single sequences in human, mouse (and other mammalian species, not shown), grey short-tailed opossum, western clawed frog, zebrafish, medaka, sea squirt, amphioxus and sea urchin. Surprisingly, two *Acss1-like* sequences were retrieved in chicken, green anole, pufferfish and green pufferfish. In stickleback, three distinct sequences were detected. One entailed a full *Acss1-like* ORF in group XVIII. The remaining two were partial and non-overlapping hits in scaffold 156 and group VI. Using intron-exon prediction tools, we found that the hit in scaffold 156 covered 2 exons with similarity to the C-terminus of *Acss1* (already annotated), while 7 exons with similarity to the N-terminus were predicted in group VI. While they could represent two different genes, we hypothesized that both sequences are different segments of the same ORF separated by a sequence gap and poor genome assembly. To test this possibility we designed primers in the predicted exons in both genomic regions to cover the sequence gap. PCR on stickleback cDNA was successful and sequencing showed the contiguity of both hits (scaffold 156 and group VI). Combining the information from direct sequencing and prediction tools on the DNA scaffolds allowed us to reconstruct the coding region of a second *Acss1* gene in this species (Supplementary Fig. 1A and B).

We next compared the full predicted amino acid sequence to examine the conservation of the typical ACS motifs (except motif IV which is absent in ACSSs) (Watkins et al., 2007) (Supplementary Table 1). A substantial degree of sequence identity was observed (Supplementary Table 1). Minor variations were detected in the sequence motif II, but not in the crucial arginine amino acid (position 18) (Watkins et al., 2007), which is conserved in all the analyzed sequences.

Various scenarios can explain the reported diversity of *Acss1* sequences. First, they can represent independent duplications of *Acss1* in birds, reptiles and some teleost species. Alternatively, we cannot exclude that the existence of two *Acss1* genes is the result of duplications (namely genome duplications), with paralogue retention and loss in distinct lineages. Finally, the extra sequence can represent a novel *Acss* subfamily distinct from the three previously described. To test these various possibilities, we started by constructing phylogenetic trees (Fig. 1). We observe two well supported clades of vertebrate sequences (bootstrap 946 and 999, Fig. 1). The group including the described human *ACSS1* gene also has orthologues in amphibians, birds and reptiles, and most teleost species (zebrafish, stickleback, pufferfish, green spotted pufferfish and Atlantic salmon). The second clade includes the newly found *Acss1* sequences in chicken, green anole, medaka, stickleback, pufferfish and green spotted pufferfish. Most importantly, both clades are outgrouped by the invertebrate chordate sequences (bootstrap 1000). The recovered *Acss1* tree topology indicates clearly that two *Acss1* clades exist in vertebrates. In light of these findings we rename the described *Acss1* sequences which groups with the human orthologue, *1A*, while the novel *Acss1* gene, we name *1B*. Finally, we note that the gene duplication which originated *1A* and *1B* isoforms predates at least the separation of teleosts and amniotes, and post dates the divergence of invertebrates.

To better delineate the *Acss1* isoform gene duplication we searched the partial genome sequence of *C. milii* (Venkatesh et al., 2007). The elephant shark belongs to the order Chimaeriformes, and is part of the Chondrichthyes, the oldest group of living jawed vertebrates. We also found two types of *Acss1* sequences in this species, but given the current coverage of the genome these are not complete. Thus, we were unable to perform phylogenetics, an essential proof of orthology. However, amino acid sequence alignment of the two partial sequences clearly suggests the presence of *1A* and *1B* paralogues in *C. milii* (Supplementary Fig. 2). Thus, we can safely argue that the duplication which originated *Acss1A* and *Acss1B* dates to pre-gnathostome origin.

### *3.2. Genomic loci of Acss1A/B in vertebrate classes indicate paralogue retention and loss*

The phylogenetic analysis indicates that *Acss1* duplicated in the vertebrate lineage, probably before the divergence of gnathostomes. Consequently, the absence of the *Acss1A* or *Acss1B* paralogues in the investigated species can only be accounted for by independent episodes of gene loss or poor genome coverage in the investigated species. To distinguish between these alternatives we have inspected the gene *loci* of *1A* and *1B* isoforms (Fig. 2). The human *ACSS1A* maps to the p arm of chromosome 20, being flanked by *TMEM90B*, *CST7*, *C20ORF3*, *VSX1*, *ENTPD6* and *PYGB*. In chicken, *Tmem90B* and *Cst7* are also found in close proximity (Fig. 2A). Overall, the *HsaACSS1A* gene *locus* content is similar to that found in green anole, western clawed frog, stickleback, green spotted pufferfish and pufferfish. However, we find three other distinct situations. None of the human *Acss1A* flanking genes are found in the proximity of the zebrafish gene. Nevertheless, *Vsx1*, *Entpd6* and *PygB* map approximately 5 Mb upstream in the same chromosome (Dr17), hinting at a possible *locus* rearrangement (not shown). Finally, we inspected the apparent absence of *Acss1A* in medaka. We used the green spotted puffer *Acss1A locus* composition to determine the arrangement of these genes in medaka and found a high conservation of gene arrangement but no intervening *Acss1A-like* sequence (Fig. 2A). However, close inspection at this genomic region allowed the identification of sequence gaps on either side of C20orf3. Therefore, we cannot conclude with total confidence on the absence of *Acss1A* medaka.

We next compared the *Acss1B* locus between the various species (Fig. 2B). The chicken *1B* gene is flanked by *Vsx2* and *Abcd4*, similar to what we find in green anole, green spotted puffer, medaka, and partially for stickleback (*Vsx2* is not found flanking *GaAcss1B*). We analyzed the location of these genes in the species where *Acss1B* was not found. Consistently, *Vsx2* and *Abcd4* are found together with no intervening gene in humans and amphibians, suggesting that *Acss1B* has been deleted (Fig. 2B), except in zebrafish where both genes localize to distinct

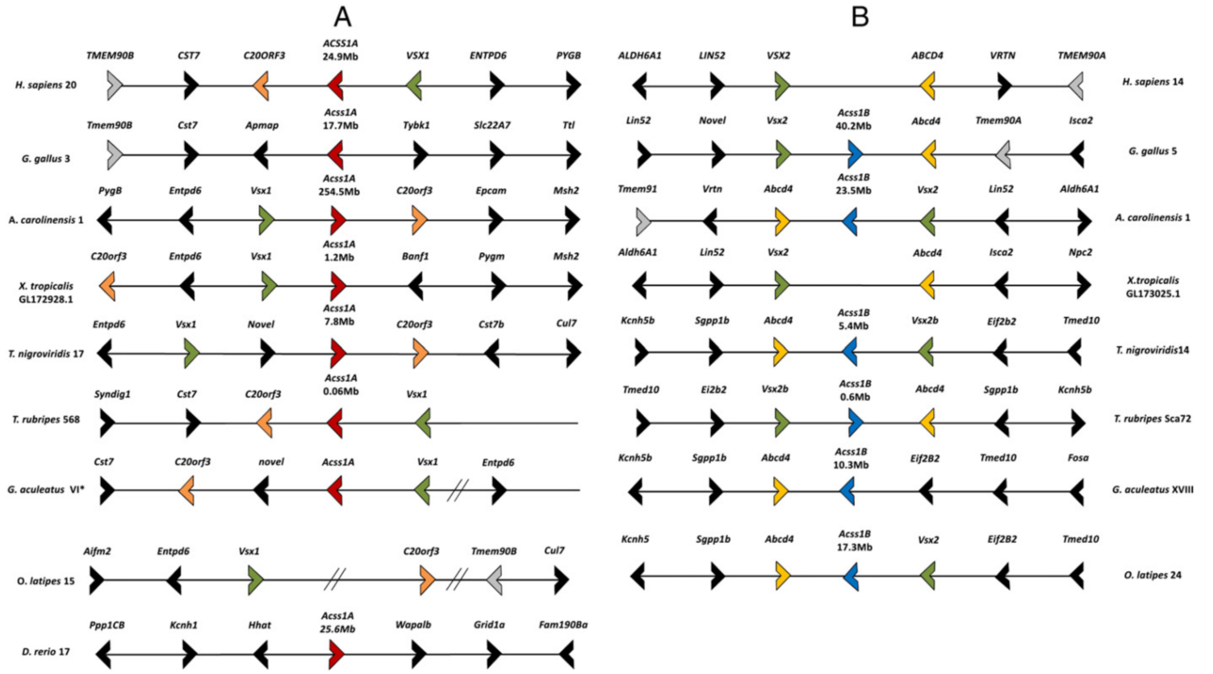


**Fig. 2.** Genomic neighborhood of the *Acss1A loci* (A) and *Acss1B* (B) in vertebrate species. Oblique dashes indicate sequence gaps. (* for details on the ACSS1A assembly see Supplementary Fig. 1A).

chromosomes (not shown). In summary, we can safely conclude that the absence of either *1A* or *1B* in the various investigated species represents cases of gene loss and not missing data (with the exception of *1A* in medaka).

From the analysis of the *Acss1loci* a curious outline emerges. We find that the set of genes surrounding both *Acss1A* and *Acss1B* (and the respective *locus* when the gene is absent) have also duplicated. At the human *Acss1A locus* in chromosome 20, we find that the majority of these genes have at least one paralogue in chromosome 14 in the region of the missing *Acss1B* gene. That is the case for *VSX1* (*VSX2*), *ENTPD6* (*ENTPD5*), *TMEM90B* (*TMEM90A*) and *PYGB* (*PYGL*) (Fig. 3). We considered whether this was an indication of the involvement of genome duplications in the generation of *Acss1* gene diversity. The so-called 2R has left genomic signatures, namely the formation of paralogons. Interestingly, *HsaAcSS1A* and the *ACSS1B*-less *locus* are part of the same paralogon (hereafter referred as linkage group (LG)) (Putnam et al., 2008). This LG, LG11, includes genomic segments in human chromosome 2, 1, and 20 (segment A), chromosome 14 and 15 (segment B), chromosome 11 (segment C) and chromosome 6 (segment D) (Fig. 3). A region in chromosome 19 (segment E) is also part of this paralogon (see Putnam et al. (2008) for details). If *ACSS1A* and *ACSS1B* are part of the LG11, then flanking duplicated genes should have paralogues mapping to expected segments of the LG. That is the case of *VSX1/2*, *ENTPD5/6*, *PYGB/L* and *TMEM90B/A*. The later two genes have a third paralogue that maps to chromosome 19 (*TMEM90*, segment E) and chromosome 11 (*PYGM*, segment C) (Fig. 3). Also, we find that the paralogue of *CST7*, *CST6*, localizes to Hsa11 (segment C). Thus, the duplication pattern observed in the genes which flank *Acss1A/B* supports the hypothesis that this was a 1R/2R-generated event. A final test of this hypothesis was the analysis of the genomic region where the amphioxus *Acss1* gene resides. If indeed the *Acss1A/B* genomic regions were structured by 1R/2R, then the amphioxus *Acss1* gene should be flanked by a set of genes that in humans localize to any of the LG11 segment regions, even if microsynteny is not conserved. We find that the genes in close proximity to *BfAcss1* in scaffold_295 are single copy in humans and localize to chromosome 14 close to *VSX2* (*FCF1*, *KIAA0317*, and *ADCK1*), and a fourth that maps to chromosome 19 (*BCKDHA*). Thus, the analysis of the *B. floridae Acss1* genomic region fully supports the hypothesis that the *Acss1* vertebrate paralogues have duplicated following 1R/2R, at the stem of vertebrate evolution.

Teleosts have undergone a lineage specific genome duplication (3R) after the divergence from other vertebrates (Jaillon et al., 2004). In the case of the *Acss1* gene subfamily we find no evidence for the retention of 3R specific duplicates in the analyzed teleost species. However, signs of 3R duplication are found in the flanking gene families of the *Acss1 loci*. For example, the *Cst7* gene which maps close *Acss1A*, has a 3R paralogue mapping at a different chromosome in *T. nigroviridis*. A similar pattern is found with gene families flanking *Acss1B* (e.g. *Vsx2*). Thus, we can safely conclude *Acss1A* and *Acss1B* duplicated in teleosts following 3R but gene loss kept single copies of both genes.

### 3.3. Evolutionary history of the Acss1 subfamily

In this work we set-out to investigate the evolutionary history of *Acss1*, a pivotal enzyme in energy metabolism. In mammals the *Acss1* subfamily is single copy (Watkins et al., 2007). Through genome database mining we investigated the portfolio of *Acss1* genes in additional vertebrate classes. A single *Acss1* gene was retrieved in some of the analyzed species. However, a previously unreported *Acss1-like* gene was found, that robustly grouped in the *Acss1* phylogenetic clade. We name the described *Acss1* genes as *Acss1A* and the newly found isoform as *Acss1B*. The new gene has a surprisingly complex phylogenetic distribution (Fig. 4). *Acss1B* is found exclusively in some teleosts, sauropsids and chondricthyians. *Acss1A* is present in all of the examined species, except probably medaka which preserves *Acss1B*. Considering the species distribution of *1A* and *1B* sequences, we hypothesized the involvement of genome duplications. The expansion of gene repertoire through genome duplications at the base of vertebrate evolution, the so-called 2R, has now been firmly demonstrated with the sequencing of the Florida lancelet genome (Putnam et al., 2008). Two complete genome duplications took place after the divergence of invertebrate chordates and probably before the speciation of agnathans (Kuraku et al., 2009). Our analysis shows confidently that from a single gene present in the ancestor of chordates, the *Acss1* gene collection expanded in the process of genome duplications in vertebrate ancestry, before gnathostome speciation (Fig. 4). Thus, we propose that the most parsimonious explanation for the scattered distribution of *Acss1* isoforms, involved independent events of gene loss and retention (Fig. 4), such other examples documented in vertebrates (Furlong and Graham, 2005; McGonnell et al., 2011; Mulley and Holland,

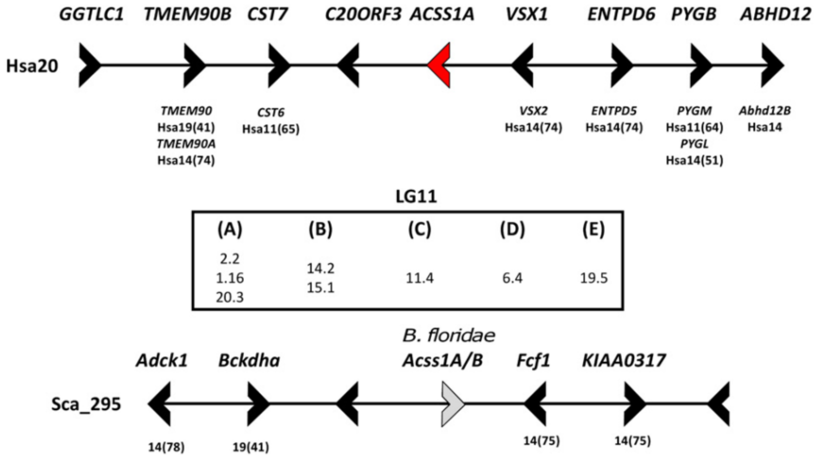


**Fig. 3.** Synteny and paralogy around the *ACSS1* gene family in *H. sapiens* (Hsa) and *B. floridae*. Human neighbouring genes with multiple paralogues have their mapping position indicated. The constitution of LG11 as defined by Putnam et al. (2008) with genomic regions from chromosome 1, 2, and 20 (A), chromosome 14 and 15 (B), chromosome 11 (C), chromosome 6 (D) and chromosome 19 (E). The orthologue location in the human genome of the gene families flanking the *B. floridae Acss1A/B* is shown for each ORF. ORFs without any indication have no clear human orthologues.

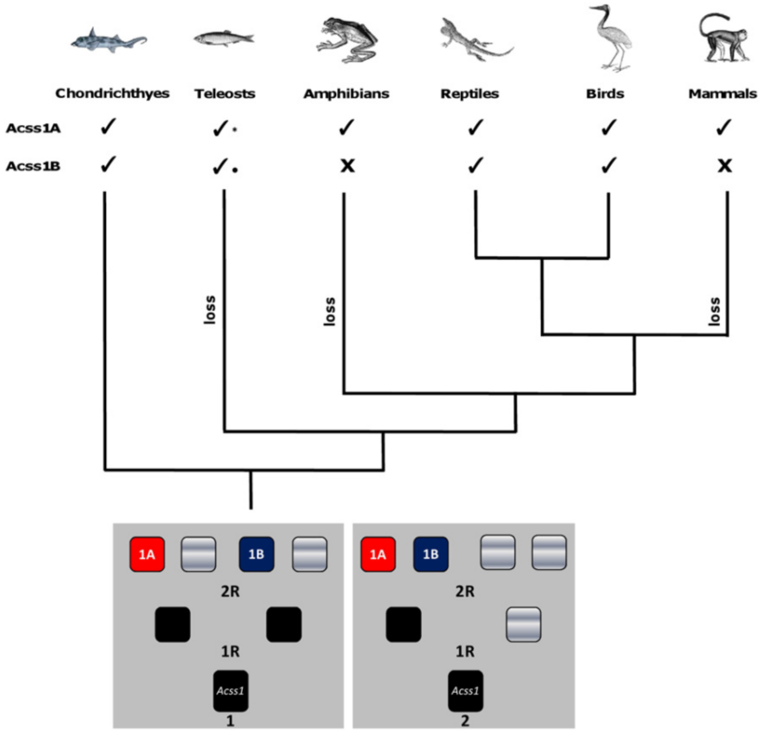


**Fig. 4.** A—Evolutionary model of *Acss1* genes in vertebrates. Scenario 1 contemplates two loss events, both after 2R, while scenario 2 implies a single loss event after 1R. Light grey boxes specify loss. Independent gene losses have taken place in distinct classes (indicated by X); * not present in medaka; ● not present in zebrafish.

2010), namely in gene families involved in fatty acid metabolism (Castro et al., 2011; Evans et al., 2008). Two alternative routes can explain the appearance of the two paralogues in gnathostome ancestry: a single ancestral *Acss1* gene duplicated to yield four genes after 2R with two being lost (scenario 1 Fig. 4); or after 1R, one loss event took place, and the second 1R gene duplicated to generate *Acss1A* and *Acss1B* (scenario 2 Fig. 4). Additional loss events can be mapped in vertebrate lineages. We find the loss of *1B* in zebrafish and the potential loss of *Acss1A* in medaka. *Acss1B* was also lost in amphibian clade, maintained in sauropsids and lost again in mammalian ancestry (Fig. 4).

Why exactly both *Acss1* isoforms have been preserved in some species but not in others is intriguing. To address this issue, we have examined the transcription profiles of *Acss1* genes in human, stickleback and zebrafish (Fig. 5). The human *ACSS1A* shows the highest transcription in testis, brain, blood and intestine (Fig. 5A). The heart, muscle and kidney also display significant levels of transcription, while no transcripts were detected in the liver and the spleen. We next determined the transcription of *Acss1A* in zebrafish through RT-PCR in adult tissues (Fig. 5B). The single copy *Acss1* from zebrafish is expressed in all the tested tissues, except the eye. Higher levels were noted in the heart, kidney, spleen and the gonads. In stickleback *Acss1A* is found in most tissues; except the

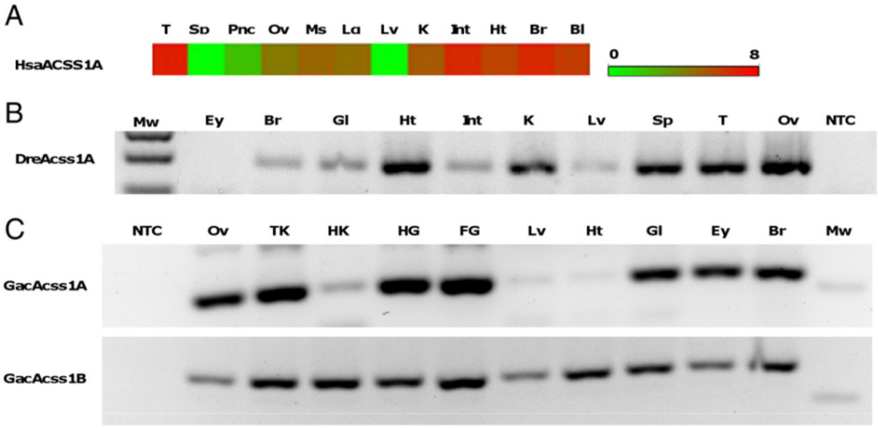


**Fig. 5.** A—Heat map of the transcription of *HsACSS1* in specified tissues represented in log2 of EST counts by color coding, green (low) yellow (medium) red (high) and white indicates no data available. B—gene transcription of *Acss1A* in *D. rerio*. C—gene transcription of *Acss1A* and *Acss1B* in *G. aculeatus*. Pnc—pancreas, Ms—muscle, Lg—lung, Ey—Eye, Bl—blood, Br—brain, Gl—gill, Int—intestine, Lv—liver, K—kidney, Sp—spleen, Ov—ovary, T—testis, Ht—heart, TK—trunk kidney, HK—head kidney, HG—hind gut, FG foregut, NTC-negative control and Mw-molecular weight marker.

heart and liver were only vestigial transcription is detected (Fig. 5C). In contrast, *GacAcss1B* is observed in all of the tissues (Fig. 5C).

The overall *Acss1A* expression data gathered from humans and zebrafish shows a high similarity to that originally described for the murine orthologue (Fujino et al., 2001). In this species the heart, kidney and skeletal muscle are the major expressing organs, with no mRNA being detected in the liver (Fujino et al., 2001). This is analogous to what we find in humans and zebrafish but not in stickleback. For example, the *1A* isoform in stickleback is not transcribed in the heart, in contrast to *1B*. Non-quantitative transcription data from chicken, where the two *Acss1* paraloques are also present, corroborates the findings in stickleback (not shown). Though we did not undertake an exhaustive transcription analysis, both stickleback paralogues apparently accommodate the full expression profile observed for the single copy *Acss1A* in humans and zebrafish.

The retention of two *Acss1* paralogues in various lineages indicates that both genes evolved distinct roles after duplication. Whether this involved the acquisition of novel functions (neofunctionalization) and/or partitioning of ancestral functions between gene copies (subfunctionalization) (Force et al., 1999), is at present unclear. Nevertheless, the independent loss of *Acss1B* in zebrafish, amphibians and mammals suggests that some degree of functional overlap between paralogues was kept. In that case, the loss of *1B* (and probably *1A* in medaka) could be compensated by the remaining isoform. Alternatively, the preservation of *Acss1B* in some species could be related with the acquisition of new functions. A more detailed expression analysis and the inclusion of other informative lineages, such as cartilaginous fish, should provide clarification on the evolution of *Acss1* function.

## 4. Conclusions

Overall, the results presented here show a novel insight into the evolution of the *Acss1* gene subfamily. We show that the diversity of isoforms is broader than anticipated, with the identification of two paralogues, *Acss1A* and *Acss1B*. The reported findings illustrate a previously unacknowledged case of paralogue retention and independent loss in vertebrate classes, whose functional consequences in energy metabolism are as yet unknown.

Supplementary data to this article can be found online at doi:10.1016/j.gene.2012.01.013.

## Acknowledgments

This research was financed by the Projects PTDC/MAR/68106/2006 and PTDC/MAR/68885/2006. The suggestions provided by three anonymous reviewers are also acknowledged.

## References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Boardman, P.E., et al., 2002. A comprehensive collection of chicken cDNAs. Curr. Biol. 12, 1965–1969.

Castro, L.F., Wilson, J.M., Gonçalves, O., Galante-Oliveira, S., Rocha, E., Cunha, I., 2011. The evolutionary history of the stearoyl-CoA desaturase gene family in vertebrates. BMC Evol. Biol. 11, 132.

Evans, H., et al., 2008. Ancient and modern duplication events and the evolution of stearoyl-CoA desaturases in teleost fishes. Physiol. Genomics 35, 18–29.

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, 1531–1545.

Fujino, T., Kondo, J., Ishikawa, M., Morikawa, K., Yamamoto, T.T., 2001. Acetyl-CoA synthetase 2, a mitochondrial matrix enzyme involved in the oxidation of acetate. J. Biol. Chem. 276, 11420–11426.

Furlong, R.F., Graham, A., 2005. Vertebrate neurogenin evolution: long-term maintenance of redundant duplicates. Dev. Genes Evol. 215, 639–644.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.

Ikeda, Y., et al., 2001. Transcriptional regulation of the murine acetyl-CoA synthetase 1 gene through multiple clustered binding sites for sterol regulatory element-binding proteins and a single neighboring site for Sp1. J. Biol. Chem. 276, 34259–34269.

Jaillon, O., et al., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946–957.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9, 286–298.

Kuraku, S., Meyer, A., Kuratani, S., 2009. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol. Biol. Evol. 26, 47–59.

McGonnell, I.M., et al., 2011. Evolution of the Alx homeobox gene family: parallel retention and independent loss of the vertebrate Alx3 gene. Evol. Dev. 13, 343–351.

Mulley, J.F., Holland, P.W., 2010. Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. Mol. Biol. Evol. 27, 2386–2391.

Orchard, S., Anderson, J.W., 1996. Substrate specificity of the short chain fattyacylcenzyme A synthetase of Pinus radiate. Pytochemistry 41, 1465–1472.

Pavlidis, P., Noble, W.S., 2003. Matrix2png: a utility for visualizing matrix data. Bioinformatics 19, 295–296.

Pérez-Chacón, G., Astudillo, A.M., Balgoma, D., Balboa, M.A., Balsinde, J., 2009. Control of free arachidonic acid levels by phospholipases A2 and lysophospholipid acyltransferases. Biochim. Biophys. Acta 1791, 1103–1113.

Putnam, N.H., et al., 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064–1071.

Rozen, S., Skaletsky, H.J., 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S., Misener, S. (Eds.), Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, pp. 365–386.

Sakakibara, I., et al., 2009. Fasting-induced hypothermia and reduced energy production in mice lacking acetyl-CoA synthetase 2. Cell Metab. 9, 191–202.

Starai, V.J., Escalante-Semerena, J.C., 2004. Acetyl-coenzyme A synthetase (AMP forming). Cell. Mol. Life Sci. 61, 2020–2030.

Venkatesh, B., et al., 2007. Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. PLoS Biol. 5, e101.

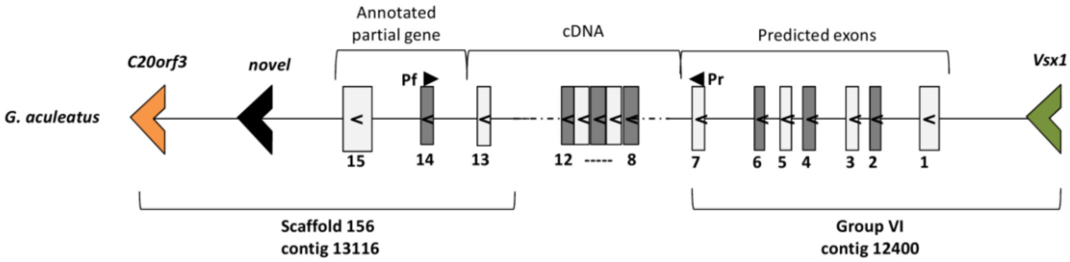
Watkins, P.A., Maiguel, D., Jia, Z., Pevsner, J., 2007. Evidence for 26 distinct acyl-coenzyme A synthetase genes in the human genome. J. Lipid Res. 48, 2736–2750.

Wheeler, D.L., et al., 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 35, D5–D12.

## Supplementary Material

### Supplementary figure 1

**(A)**



**(B)**

```
>scaffold_156 dna:scaffold
scaffold:BROADS1:scaffold_156:133778:170221:1
GTTAAACTCGGAAGCCAGGTGTCTCAACTCTTCTCCACGAACACGCCACCGAGGCTAGGCCTCCACAGCCCCCGGATAAGAAGTTAA
AAATCTCCACCCCAGGCCATAAGTAGCCCCCGGGGAGATCGTCCACATTCTACAGAGGGGAGAGGTTTTTTTTCCTCACACCGTCTT
AAACAACCAACGGGTGATACCCAACCCGGATTTTCCCCACTAGTCCTTGTGTGAACGCCAATTAGTAGTTGTTTCTTTCTTTTTTTT
ATTTTGTGGAGAGTGTGTGGGGGGGGGGGGGTTCTTTCCACTGGAGGAAAAGGAAAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAA
GAAGACGACGTTGATTTAGGAAACTTTTTTTCAGTTAAAATTATATGGTTTGTGTGCGCAGCCAGATGATCCAGTGGAAGGAGGCCA
AACGAAACCACAAACTCCTCCAGACTGGAACACTGAGACTCACAGCTGCTCCCTCATGTTTGAGGAATGAAGCTTCATAAACACTGG
ATGAGGTCATATGATCCTTTTAAAGGAGGACCAAGGGCCAGGTCCACATCGGTGATGAGGAGGAGGACCACGGACCGGGTCCACATC
GGTGATGAGGAGGAGGACCACGGACCGGGTCCACATCGGTAATGGGTCGATCCCTTGTGATCACTTCACACGGGCACACAAAGAACT
CCACCCGTAATACTTGAATACAGGTTCGAGGGAAATCTTCTCCAGCTACATGGTTCCGTTAGTCCAGCTACATGGTTAGCACACGCT
AACAGGTGTGATGCCAAACCCAAAGGCCAACATGGCCCTCCACTATATCCTCCAGCATCTGGACCCCCCAGGAACCTACACCAGGAT
CCTGTTTGTGGACTTCCGTCTCTGCTGCAGGACAAACTCTCCCAGCTGCACATGCCTGACTCCACCTGCAAGTGGATCACAGACTTC
CTGTCTGACAGGAAGCAGCACGTGAAGCTGGGGAAACATGTCTCAGCCTCCCAGACCATCAGCACCGGTTCCCCCCGAGGCTGCGTT
CTCTCTCCTCTGCTCTTCTCCCTGTACACCAACAGCTGCACCTCCAACCACCAGTCCGTCAAGCTCCTGAAGTTTGTGGATGACACC
ACCCTCATTGGACTGATCTCTTGTGGGGAAGAAGGCTCAGCAGAGGTTGTTCTGAAGAAATTCAACCTGCCAAAGACGATGATGGTC
CACTTCTACATGCCATCATGGAGTCCATCATCTGCTCCTCCATCAGCGTCTGGTACGCTGCAGCCACAACCAAGGACAAGGGCAGGC
TGCAGCGTGTCCTCCGCTCTGCAGAGAGGCTGATCGGCTGCAATCTGCCGTCCCTGCAGGACTTGTTCGCTTCTAGGACCCTGAAGC
AGCTAAAAAGATGGTAGCCGACCCCTCTCACCCCGGACAAAACCTGTTTGTGCCCCTTCCATCTGGCAGGAGGCTGAGGGGACTAAG
ACTAAGGACTAAGACCTCCCGCCACACCAACAGTTTCTTCCTGTCGGCAGTCGGGCTCATCAACAGAGCCGGTCCCCCCCTGACTGA
CTCTGACATCGCACCAGTCACTCCCTTCACACTGCACACGGACCCTGTCTTGTTGTCCAACTGTCTGCTGCTACGCACCAACCGCCC
GGTCAAATTCCTTGTATGTCTGACATATTTGGTATGAATGGACCTGAGAAAGTGAGGCTCACCTGGGGCTGCCGTGGAGTGACGGCG
TTGAGCTGGCCTTGCGGAAGCCGTCCCTGAGGACGGGGCGACGGCCTGACTCCTTCTCCGACCTGGAGCGGACCGGAGGGCCTGGCT
CGTTGTCTTTGGAGCTGCCTCGGAGCCCCTCCCCCGAGGTGGCGAGCGCCCCGTTGGAGCTAGACGTCTCCGAGGGGGGGGCGTGCG
GGTCCGGGTGGGCCAGAGGGGCAAAGCTCAGCTCGATGATTTGCTTGGCGCTCTCGCAGATCTGACTCTGAACACACATACAGACAC
GCACACGTTATCTTCGGCCACACAGGGCGAGTGGAGGCAGGCCCCTCCCCCTCAGCCCCGTACCCTCCCTACCTGGATCTCAGGCGT
GAGGGTGGGCAGGTCGAAGGTGAACACCGAGGTGGACCCAGAGGTCAGGAGGTCCACGCTGTCCAGGCTGCTGTAGGAGGTGCCTTT
GGCCCGGTGAGACTCCATGATGCTCTCGAAGTGGCGGCTGAACGAGTCCAGGTTGCTCTCAGAGGGGCGGCGGGGGCCGCCGATGGA
GATCGGAGACTTGAGGCCGTCGCAGGAGTCTGAGGTCACCAGCGACCTGTGGGAGGCAGAGGAAACTCTTCAGCCTCAGCTCCAGGA
```

```
CGTCAGAGAAGCCACCACAGCATTGATCGATACTAATCCTTCTGATCCTGAGTCATTAAACAGACATGATGACGTGTTTATCATCAA
TGAGGTGGTTATTAGCTACATTACTTTTTAACTTTATTGATTCTTAACTCCATAAAAGGACCAAACTGAGACCAGGACGATTAAGCA
AACGGTTAAAATGCATCGAGGGATCAGCAGAAGAACGTGACACTTATTGTTTTACGCTGTTCATTCTGTACACATGACATCCATCGC
AGTCGGTCTTCCTGACAGAAGGATCCTCCTCTGTCTCCCTAAGGTATCTTTCTTGTTTGGGGAGTTTTTCCTGTGCCGATGTGAGGG
TTTCGGGACAGTTAAACCCTCTGAGGCAGATTTGCGATGTTGGGCTGTACAAAATATAATGAAGTGAACTTCCTGTTCTCTCGGGCC
TTATGGGAGCTGTAGTTTGAGCCGGTGGATCTCACCCCTTCAGCAGCAGACTGAAGACCTCCTCTTCTTCGTTGGTGGTCCTTTGTC
TCTGTTGATCTCCGAGCATCCGGACAGTGGCCCAGCTCACCACTCCTTCCTCCTCCTCCTCTTCTTCTCCATCCTCCACGTCTC
CACCCGTCTGGGGTTTAGGGGAGGCGTTGGCCGGGCTGCAGCTAGAGTTTTCCAGGACGCTGTGGTTGCTCCAAAAGGCCGAGTCCA
CCTCCTCGTCTGCCGCCTCAGCTTCCGCCCGCATCACCTGCACCACAAAGACGACCGTTGTCAGCAAGAAGATGCCATCATAGAGGG
AGAAGTGGGTCACGTGACCTCAGCCCACCTCGTCCAGGTCCGTCTCCCTGTAGCACCGCAGCGGCACCGCGTCCAAGTCAGTCTCCG
AGTACTGGAGAGTGTGGCGGACGATGCCTGTCCTGGGGGTGGAGCCTTGGGCGTCAGGGCTCACTGGCTCCACCACGATCTGCCGCA
CCTGGCGATGCAATGGTCCTGGACCCTCGCTGCCCCGCCCCCTCCTTTGATTGGCTGGTTTCAGATGGAGGGAGGAGGGAGGGGTGG
GGTTCGGCAAACTGCTGGCTGGAGGCGTCGCAGATGGAGCTTCTGCAGGGGGAAAGAGTTTTATGACCTAAAGATGTCCCAAAGAAT
ATACATTTTATATATATTATATATATATATATATATCAAGAAATTCTGTAAAATTAATCATTTTGTATGAGAACAGTTGGAACATTT
TCTTGCTCATTCATCTGGTCATATACACCGCCAAGACAAATTCCATGTATGTTTGGCATATTTTGGTAATAAATGTTTCCTGATTCC
TGATATGTACTGGTGTGGTTAGTTGCATGTAGTACCTTGTACTGTGAGTCTAAGTATGAATCAAGACGTATTTAAACAGTTATAAAC
CAATGCCTGGAGAGGATGCTGGTACCCAACAGGAAGTTGTAGTAGTAGTAACACTTGCACACCAAGACCTGTTGAGCTGTATTAAGC
TTCAGCTAATTGGCTTAGAGGAAGACTGGAAGGTAACGAAGCTGCTTGTTTACCGAGAGGAGACAAGATGTCACACCGGGACATCTA
CACATAGTACTGATACCGATCAATGCTGATTTGGCTTTTGCCGTGTCTTATTGACTATGTACATATTCATAAAAAGTAAACAACCTC
TCTCCCCGAACAGGAAGTCCCACTTGGCTCTGGCGATCTTCTGCGAGCGCGACGATGAGGAGCCTTCTGACTGGAGGGTGGAGAGAG
GGAAGCAACTCGATCAATAACCAGAGTTATTGATTCACCAATGCAGTGTTTCCCCAACCATTATAACACCCCCCACCAACCCCTAAA
AAATGTGCAATAAGTAAAAGTAAATAATATGTTAAAATTTGTATCATTTTTCTTGTTTCTTTCCATGCACAGCTGCACTGACATGAA
TATGGCACAAACATGTCTGTGTGCTCGAGTGACTGAACCCCCCGAAGTTCCAAACCCAGAGAAAACACTTCTCCTCAGCTTCATTTT
TAAAGGGAATACAATCTTACGCGAGTCCATTTTTTGAAGCTGCCAAACTCAAAAGCAAAGAAGCAAAAGTAAAATAGAAAGTAACTA
AAAAGAAACAGTCTTGTTAAAATTATTTATTTACTTATTTGGGGCCTAATGGTGCGTTCACCCCAAAAGCGAAGTGAATTTTTGTCC
CGCTTTACTCGCACAAGTTTACTCGCGCAAGTATTTTGCTTCATTCGTGTACAATTTTTTTCGCTTCATCGCGCCAGGAAGGGGGCG
TGGCTTATCTCCGTTACGTGTTTCCGTAAAGGCAGTTATCGTTTCCGCTATCCACCATGGCAATAAATAACCACAATTCCTGCTCTT
TATTTGTTGTGTGAATTACACAAACGTCGGAACCGCAAACGGTGTTTATTCGCTTCGCTCGTTTCACTTTTTTGCGTCCCACCCGCC
GCGTCTTTCACGTCATTTGCGCCGCCTGTTTAAAATCCGCCTCATTCACGCCGCGTCTGTGCATTGACGTTACATGGAATCTACTCG
CGCGAATAATTTGCTTTTGGTGTGAACGCACCATAAGGCTCATGGGAAATATAGTCTGTTGTATTGATAAGGTCGATCACCATGACA
ACAGAAACTCACCGGTGGGGAAGGATCTGGATTTCGGATTTATTTTCCGGTTTACTGTCATCGAGGTAAATGTGTCAATGAGGGAAG
AAAGACACCCTCACGCTGCACAAAACACACAAACCTACACACCAATCCCCCCCGACCCTCCGACCTGCTTCCCTGAGCCCGTGTCT
GCCATGTTGTGGCTGGAATGGCTCGAGGTTTTGAGGGACGGGCTCCCGTTTCTTTCGGCCGTGCTGCCGGTGACCCCACTGGAGCTC
CGTGAGGCGGAGCCAGCGGGGTCACTGTGGTCCTGAAGCTCCTCCCTCTCCCCGTCGCCTCTTCTGACCTCCGTGCGGCGCTCTTCT
TTGCCTAGAGCATACCCTGCGGGAGTCAACCATCCAAAGCAGTCTTCGCCGAATCCCTCCCTGTCCCGGAACGGTCTGACGTCCGAC
CCCCACCGGCTGGCGTGGACAGCCGACCTGGAGGGTTCAGGGGGCTCCTCAGGTCCGTGGCATAGGCGATGATGCCGCTCCGTCGAT
GGAGATGAAGAGGTTGGGTACAGCGGACTAGGCAAGGTTGGAGACGCCGGTGATGATGGGCCTTTCCTGGTCAGAGGAGGCAAGGCG
CGCCCCGGTCCGGCTCGAGGCGTTTTCCATCGAAGGTCCGGGTCATCGGAGTCTCTCAGGAAGGAGACATTGCTGGTCTCTGACCTC
GGTGACCCGGCCTGGGAGGACGCCCCTGGAGAGGGTGTCCGCCTCTCGGCGTACGTGGTAGACAGCTTGGTGGCGTCCACTGCTAAC
TTACAAGCCACTTCAGAATTTATGGTTGGGGAGATGGTCCTCCTGCCGTAGGTGGGGCTGACCGAGTCGCGGGGGTGCCTGCTGCAC
ACGTCACCACAACTACTGCAGGGTATACTTTGTGTGCTATTGTGGGTCATGTTGGTGGTGCTAGTACAATGGCTGGCTAGACTGGAA
```

```
TAGTTGCCGGTGCAGAACCGGCCAGTCGGGACCTCGTGCTGGATGTCCGTGGGCCTCCCCGCCGGTAATGCTGGGTGACACTTTCTG
GACAAGCTCGGGCTCTGCCCCCCCCCCCGGAAGCCCCGAGAGCAGAGAGACGAGGTGCAGGCTAACGAGTTGGAGGGGCAGCCGGTC
TGAGAACACAGGCGGTTGGAGGCGGGGCTTCTGGGTAGTCCATTCGTCGATCCACGGCAGGCTACGCTGGCCGGCAGGGTGAAGGCG
CTGCGGGGCAGGACGGGGGACCCCCCCCAAGACCGGCAGCGGGGGGAGGGGTGCTGGTAGTGCCGCGGCAGGGTTGGGCTGTAGTTC
CTCGGGTTGTTTTCGCCAAATCTGCGCCTGAGTCCCGGGGATCCGAATTCCTCCAGTGCCCTGCAGGTGGCGTCTCTGGCCATTGGG
TCCGGGCTCGCCCTCTGCAGGTAAGACAGGCCTTGTGGGGAGCCGCTCCTCCGCGAGGGCGGTAAGGGGTCCCCGTGCCCCGGCATT
GCGCCGTACCGCACCGGGTCACTCAACCTCTTTGGGAGCCGGCGCCCCTCTTCCAGCCAGTGGAGAGGACTGCCCCGCACGCTTTGG
TCGCCCCCTCCGCCGCAAACATTAGCTTTCTCGATGTAGCCAAAAGTGACCACGGACGTCGTCCCCTCCTCGCCCGCGACACCCGTG
ACCTTGAAGGACCCACGGTGGCCGGTGGGCAGGAGGGGTAGAGAAGACGTAGCAGACGGGGAGATGAGCGTCTTTTCCCGGGAGGGA
GCCCAATCGGAGGGACAGGTGCATCCCAAGTGCAACCCACTCTGGCCCAGGACGGTTGCTCCGGGTGCAAGGACCTGGGGCAACTGG
CTCAAACCCTCGTGGGACAGATCTTGCCTGTGGAACTGGGTGGGACACCCTGTGTCGTCTGGGCTCTGGAGTTGGAAGCTGACGGAG
TGTCTGGAGGAAGACTTTGAGGGGGACGAGGGGTCAGGGGGTGGAGCCTGGCGGCAGTCGGGGCTGTGGGACTGAGGGATGACGTTG
GAGCACTGAAGCATCTGGTGGGCGGTCCCGTCATCGCTTCCTCCGGGCACCTGATGTTCCTCCTCGGAGACGGAACGGACCTCGACG
TACAGGTGGAGAACTTTACCTGCCTGGGACATCGCGGCCCGTCAAGACACTCAAAGTCCCTTCCTGTAACTCGGTAAACCGTGTGAA
GAAGGTCATCAGCGTCCCTTTGACCTGCAACAGGAAGCAGAGGTGTAGACAGAGGTTATTCTACACAGTTCAAAATGTGCAGATGTG
CAGACTGAACATCTGGACAGCAGAGGATTCTGGATTTCACTCTGTTCAAGGTCTATTACACATGCAGGCGTCCAGAGCGGGACCAGC
AGGAGGACAGTTCGAATTGTTGATTAAGTCTCATGAACAAATATTTTAAAAGATAGTAAACAATAGCAATTTAGATGCATTCTACAT
GTTTTCCTTCTTCTTCATTAAAACATTTTAATATTACTATTATTGTTATATATTATATGTATATTAAACTTCCATTTTTGTCTAAGA
TTCCTTGAGTGTTATTACATTTTTGTATTTTGATTTTTCAGTAAAGAAACAACAAAAACATTTAATTTACTTAATTTAATAAGTTGA
CCTATTATTTGATTTATTATTGGGAGAAACTAATCAAACTAATTAAAGTAAATCCTTTCATCTTATAAACTGTTGGGTTTGTGAATT
GCCCCAAACGTCAGATCAATCATCTTATTTGTTGGTAATCGATCGGCTTTTATCTGAACTCAGTTTAAAGATTGCGTCACTAAATTA
TTACACATGAATTTAAACCTCAGAGGTACGGAGCTGTGAGACACTGTGTGTGACTGATAAGCACACAGTTAGCTCTACTTTTCCTGG
ACTCGGCTAATAGGATTCCCTCCCGCCCCCTTTCCCCATCCAGCCTGCACTGCTCAACACACACACACACACACACACACACAAACG
CACACACACCTGTGTGTCCCGTGTCCTCGGCGACAACGCCTCTGCAGCAACGTCCCCCGAGTTGCCGTGGTAACCTCCTTTCTCCAA
CTCCCCCTTGGTTAGCGAGGACGCCGGGGCTCTCCTCCCTCTCTGAACACTTGTTGTCTGCAGTAGAGCCGGCTCATCGTGGATTTA
CGGGCTCCACTCGGCTTGTCCTCCTCTTGTGACGCCTCCTCCTCTTCCTCTTCCGATCTCCTTCAATCAGCTGACATAGCCTCCCCT
TCTTCTGGGCAAGCTGCCAGCCAATCAGGTCCTCCGCAGCCTGCAGACCTTTCTTCTGTTGGCCAATGGTAAGCATCTCCGACAGAT
GGCTCAGGTTATTACAGTGTGTGAGGGGGGAGGGGGGGGGTGACCTGCTGCTTTATAAAAGGACAATTTAAAGTATTAGAGGAAAGA
CACAGACTGCACATTACTCTGTATGGCCACACCCCCGGCAGTGATGTCACAATCCACCCACATCCACACCTGTGCATGGTAAGTTCA
AGTGTCTGTAATGACAGTAGTAGAGTCTAGTGTCTGTGGTGAGTTCAGGTGTCTGTAATGGCAGTAGAAGAGTCAGGGGAGTGTGGT
AGGTTCAGGTGATGCAAGGTTCATGCTGCTTTCAGGTGCAGTCGTCCACTCTTTCACCAGCAAGAGAAAATTTTGAACAAGAAGTTT
GAGAGTAAAAGTTAAAGCGCAAAATACACCTTTTCTGGTTCAGATTTTTCTGGTTTCCGTCCCTCCCTGTGGACTACGGGCACCGTT
CAGAGACCTCATTCTACTGCACCAGAGGTCTCACAGGTTGACCCTGTCAGTCCTGGTTCTTTTGGGCCTCCCTTCCCTCCTGGGAGG
AGCTCTGCAGAGAACGTAGGTCTGTCCACGGTGGAGCGGTCTCTGCTAGGGGATCCGAGTCGGCCCACGGTGGACTGGTCCCTGCTG
GGAGATCTCGGTCTGTCCATGGTGGATTGGTCTCTGTGGAGGATCTGGGTCCGTGTAACAACCTCCTTGACCCCCACCAGTCAGGCT
TCAAAGCAGGCCATTCCACAGCGACTGCTCTCCTCGCTGTCTCAGAGCGACTCCACGCTGCTAGAGCCGCTTCTCTCAGGAATCAGG
AAACATTTATTACCATAATATATCAGACATACAAGGAATTTGACTTGGCGGTTGGTGCATAACAGCAGACAGTACGACAATGGACAA
TCAGACAACATAGTGCAAAATATATATATATATATATATATATATATATACACATATATATATAATGCTATGGGTTAGTTTGAAAAT
AAAGGAATAAAAGTTTAAAAAGTTAAAAATATTTAAAAATTAAAAATAAAGTTCCCCAGCAAACAGAAAGTGACAAGTGACGTGTGC
GGTGTCAAAGAAAGTAACCGGTGGGATGTCAAAGTCAGTCAGTGGGGGACCGGGCTCTGTGGAGCCCGACTGCCGACGGTAAGAAAC
TGAAACTGGTGTGGCGGGAGGTCTTTGTCCTGGTGGACCTCAGCCTCCTGCCAGATGGAAGGGGCACAAACAGGTTTTGTCCGGGGT
```

```
GAGAGGGGTCGGCTACCATCTTTTTAGCTCGCTTCAGGGTCCTAGAAGCGAACAAGTCCTGCAGGGACGGCAGATTGCAGCCGATCA
GCCTCTCTGCAGAGCGGAGGACATGCTGCAGCCTGCCCTTGTCCTTGGCTGTGGCTGCAGCGTACCAGACGGTGATGGAGGAGCAGA
GGACGGACTCCATGATGGCCGTGTAGAAGTGGACCATCATCGTCTTTGGCAGGTTGAGTTTCTTCAGCTGCCTCAGGAAGAACAACC
TCTGCTGAGCCTTCTTGGTGATGGAGCTGATGTTCAGCTCCCACTTGAGGTACTGGGTGATGATGGAGCCCAGCAAACGGAAGGAAT
CCACAATAGTGACGGGGGGGTCACACAGGGTGATGGGGGAGGGTGGGGCTCTGTTCTTCCTGAAATCCACAACCATCTCCACTGTCT
TTAGAGCGTTGAGCTCCAGGTTGTTCTGGCTGCACCACGACACCAGATGGTCAGACTCCACCTGTAGGCGGACTCGTCCCCACCAGA
GATCAGTTCAATGAGGGTGGTGTCATCCGCAAACTTGACTGGAGGTGCAGCTGTTGGTGTACAGGGAGAAGAGCAGAGGGGAAAGAA
CGTAGCCTTGGGGGGAATCGGTGCTGATGGTCTGGGAGGCTGAGTTTTCTGAGTTCCGCTGTAAACCAGGTTTGTCGTTATTGTAAC
TCACCCTGGTGCATGATGGTACACAGCTGTCCTCACAGAAGCCGATGTAAGAAGTTATAGCCTTTGTGAACTCATCCAGATTGTTAG
TAGCCGTCCTGAAAACATCCCAGTCTGTGCAGTCAAAGCACGCCCAAAGATCCTCCAGCGCCTCACTGGTCCAGTTCTTGAATTGCC
TCACCACAGTTTTGCAGAGCTTTAGTTTCTGCCTGTATGCAGGAATCAGATGGACCATGACGCGGTCAAACGCAGACCAGGATGAAT
GAAGCAAACTCACGGGGGGAGTAAAGGGGTTTACAGTTGATGATGAAAGTTTCCAAAGATTCCATTCTCCTTGGTCCTCATCCTTCT
GGACCTCTCTGCTGCATTTGACACAGTAAACCACCACATCCTTGTCTCCTCCCTTCAGGAACTTGGTGTCTCAGGGTCTGCTCTCTC
CCTTCTCTCGTCCTACCTTGATGGCCGCTCCTAACGGGTAACCTGGAGAGGATCTGTGTCGGAAACTTACTAGCGGAGTTCCTCAGG
GTTCCGTCCTGGGTCCCCTCCTCTTCTCGCTCTACACCAACTCTCCAGAGAGAGTTGTCATTCGCTCGCATGGCTTCTCCTACCACA
GCTATGTCAATGACACCCAACTAGTTCTCTCCTTCCCTCCCTCGGATCTCTGCCTGTCTGACTGACGTCTCTCAGTGGATCCGCCCA
CCATCTGAAAATCAACCCCGACAAGACTGAACTACTTCTCTTTCTGGGAAAAGATTCTCCTACCCAGGACCTGACTGTTAACTTTGA
GACTCCGTGTTAACGTCCCCTTTACTGCAGGAACCTCGGCGTGACATCGACAGCCGACTCTCCCTGACTCCAACATCACTGACAACA
CCATCCTGTAGATACTCGCTCTACAACATGAGGAGAACACGTTCTCTTCTCACTCAGAAGGAGGTACTGATTCAGGCTCTTGTCATC
TCCCTCCTGGACTACTGTAACTCTCTCCTGCAGGTCTCCTGCTACCTCCATTCCACCTCTGCAGCTCATCCCGAATGCAGCAGCTCC
AGTGGTCTTTAACCTCCCTAAGTTCTCCCACACTCCTCCACTCCTCCGCTCTCTTCACTGGTACCGGTGGCTCAGCCAGTTCTAAAC
ATGGTGCTCACGTACCATGCTGTGAATGGACGGGGTCCAGCTTCCATCCAGGACCTGGTCCAACCCGACATCCCGACCCGCACTCTC
CGCTCTGCATGTGATAAACTGCTTGTTCCTCCTCACTGAGAGCAAAACACTCGACTAGATCTCCACTCTTTGCTGTCCTGCTCCTAA
ATGGTGGAAGGAGGTCTCTGAAGACATCAGGACCACAGAGAGCCTTCACATCTTCAGACTAAAGACACACCTCTTCAGACTCTACCT
CCACTAACACACTAACTAACTGTAGCACTTACATTGGACTTATAATGGTTCTTATCTACAGTTGTACATCAGCTCCCTGGATGAAAT
TGCACTTTCTTGTTTTTTTTTTCTTCTGAGTTTGTTTCCTTATGGTTGAGATGTACTTGATGTAAGTCGCTTTGGATAAAAGCGTCA
GATAAATGACATGTGATGGACTGGTCCCTGCTCAAGGATGTGGCTCTGTCCATTAGTCACTCACTATGCTCAGTGTGTGTGTGTGTG
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTATATGTCTGTGTGTTGCACCACCTCAGTTACAGGAGGCGGAGTCACGTGGGCCA
ACGACAGGAAGTCTAAAGAGGAGAGGAGGAGAGTTTATAATATATGAACTCCTGTGTTTGTGTGTTTTATTTTGATAAAGTTAAATA
TAAAAGAATCATGTGCGTGTCACAGCCGGACACATTTTTTTCATATATATAACTACAATTTGTACATTTACTCACTTATAATTATAT
ATTTATTAAGTAAAATATGTGGAATTTCTTTTAAAAAAAATATTATCAAACCATATGAAAAATAGGAATGAATGAAAAAAAGAATAT
GAAAACGAGAGAAAATTATTTTAAGCTACATCATATTACTCATACTCATCATAATACTGACGAGAGCACCAAAGAGTGATTATTACA
GTGAATTACTGATAATATGTGATTTTAAGGATCAGTAGGAACCAATGAGCTTGGAGTTGAGACAGGAAGTCAGACCGAGTTGAGGGA
CGGACTGACACACTGCTGGTTTTATATGAATCTTTACAAAAATTGAAGTATTATTGTTGTTGTTTTTTCTAAATAAAACAATCGTGT
CTTCAGTGAGACAAAAATAATATTGTTATATATTATATATATATATATTATTACAAACTGAGAATTTGGCATTTGGATTTCAGATAA
TATGTTATGGTTTCTGATTCCACAATTTAAAATATTTCCCTCATTTGCTTTGCAAACAACAATTAATTGAATTTTGAATTAACCTAT
TAAAATTCCCTTGATCAATCTATTCATTGGCCCGTCTCGATGTGCGAAATCGGTGTCTCTTCTTCCTCTAGATTAATAAAATGTATA
ACTGAGCACCTGACTTGTTGCAATTAAATAAAACAGTAAAAATATCTATTTGAGGGCGTCAACTTACCGTGATATTGTTTGGTAAAC
ACTGTCAATATTAAAGACTGATATCATCAATATATCATCAATACTGAACATAATGTGGTTATTGCAGCTTCTTTATTCTCATAAGAT
AAAAGATGTGATCAAAGAACAAAGGAACAAACCGGAGCTCTTTAATCAAAAGCTTTCGGTCTGTAGGAGCTTCTTGTTTCCAATAAT
CAGTATCGATCCGACGCGTCCCTCTCCTTCACCCTGACTGATACACACGGTCACTTCGGACCTCTTCACCAGTTTGTCTTAGCAGTT
```

```
TGAACGCTGGAACACAAGCGTAAAGCTTGCTCGAGGGCACACGGAAGAGCCGTCCTCTCCCGGTGGTTCTCGTGCGAAGCCCCCCGT
TAGGAGACGCTTTGCTCGCGGTGCACATGCTGCTAAAGGACGCACGCCGAAGGGAGGAGGAGACGCCAATCAATCGATACAGATGTT
AAAGTACAGAGGAGCGTTTATACAAACTAGCTAACATCCCCGACGGTTTACAGCGGTGCGTTTCACCAAAGGAGTTCCGCACTTACC
CCGAACGCGACGATACGTTAGCTGGCCGCTTCCTACCGTCACCCGGGGGGTCCATTCATCGGCGACGGCTTCACTACCGGGCGAACC
GACGGCCTGTTACGGGTCAAGTGACCGGGAGTAACGGCTAGCGGAGTTACGTTACATGCCGGCGGAGCTCCGCTCCGTAACTAACGT
TACTTCTTCTCGTCCTCCTCCGTGTTCTCCTCCTCCTCCTCCTCTCCTTCTCCTCCTCTCCTCTCTCTCTCGAGCACCTCTTCTCGG
TGTTGTTTTTCGGCGGGTGGTTGACATCAGTTTACGTTCTCCACGGTGACTTCGACTTCCGCATCGCTGTGGTGACGTAACGGTCAC
GTCAGAAGACGTCAGAGGGGACCGACTGTCAAATAATGAAATAAATCCTAATAATGGTAACAGCGTGAACTTATTCTGCTCAATGTC
TCTTTGTTGATTTGTATTTTAATTTGTGTATACATTATATATATATATATATACATATATAATTTATAGGATCAGTAGGAACCGAGT
TGAGGGACGGACTGACACACACACATATATATATATATATATAAATATATATATATATATATATATATATATATACTTAATAATGTA
TTTCTACACAGTGGTACTACTACAATAACATATCAGAATACTACTGCTTACACACATGGAGCCCGAGCTGAAGATGAATAAGAGTTT
CGTGTCTGTCGGAAACATCCACTCTCAGTGTGAGGTATTGACCTTCAGTTGAGTCCCTCTTCATGTGAGTGCTTTGGGTTTATAAAT
GAACACCAGGCGCAGAACAATGGAAGAGTTAATATTCATGATGATATTAATGTTGTTTTTCAGTAGAACCTTTTCACGAGTAATAAT
ACTCTAAGGCAAAATATCTCAAAGTAGAACAGTCATTCGTTCAATATTTTCGTTTTGCTGAACCGACACTTTGGTAATAGTCAGTCA
GTAAGTAATTAGTAAGAAGCACAACAGTTCAGTAGAAACTGAATAAAAACATTTAAGTTCAGGTGGAGAAAAAAACGCAGGTGTATC
ATCTAAGCCCCGCCCCCAGCGAACCGGCTTTATAATTAGCCCGGTGAGGGAGGACGAGGACAGACAGACAGAGAAGGAGACGCGGAC
CGGAGGCTCCGGGCTCATACAGGTATCAAACGTAACGCCACAATCACCATGCTGGAGGGGAACGGGACGGAGCTGCGGCTCAGCGAG
TTCGGCTTCGAGCGACGCTTCGACGAGAGAGGCGCCATCGAGTGGATGCAGGCGAACTGGTGAGCGAACAATCTGATCCTGATTAAA
TGCAGGCGAACTGGTGAGCGAACAATCTGATCCTGATTAAATGCAGGCGAACTGGTGAGCGAACAATCTGATCCCGATTAAATGCAG
GCGAACTGGTGAGCGAACAATCTGATCCTGATTAAATGCAGGCGAACTGGTGAGCGAACAATCTGATCCTGATTAAATGCAGGCGAA
CTGGTGGGCGAACACCGTGTGTTCATGTTTCACTCTTTGATTCAGTAGCTTCTTGGAGGATGAAAATCACCACACAGAAGCGGAGAA
CCTGCCATTTGGTCACTTTTCCGAATCCATAAAAATAACCGTCACTTCATAAAAAACTCATTAAACGTGCGTGAAGGCGGGTTCGTG
GGGGTTCGGCCCGGTGTGCGTTTCTGCACACACCCAGGACACCGAAACTCTCCGTTTAACAAACATTCACCGTAACCGGTTCGATCG
ATACTTTAACGCAATGCTCCACTGGCGCGATCAGAGCGCGCGGGTCATTCTGTGTTCGTGCCGTGATCGGTTCACGTGACCGGATGG
AGCTACGTTGATTATCGCGATGAGGGTGGGATAAGCGCGTTAATATGTGGGGTGGGGGGCATTGGAGAGCTCTGATAAGTCCGCTCA
TTGGATGATACAGGCGTACAGCGGGTGACGCAGAGTTACGTCACACTGATGCATCGGCGGTAAACCCGGTTATTCCAGTTAGTCCGC
ACCGTCACTCCCAGGTGATGAAGTGTTAAAGTATGTACCAGACACTCAATATATTTCACCTGACAATAAACAGAACTAAATAAATGA
GACCATATGAGACACAAATTATTTATTATAATACATTTTTATGTTTACCTGGTCAACCATAAAATACTTGTACTATAGAAGAGTTTC
ATTAGAGCGACTGATTCGTGAAAATAAATCTGTGGAGTGGGAACATCTGGTGACTCAGTGGGTGGAGCCTCGGGTGAATTCACCATT
CTTTGGTTTGTGGTTGCTAGGCAACCTAGGTGTGACTTGGCCTCAACACAAAACGCGAATGTTTCTGCTAGAGAACTTCGTTGGCTT
TGAAGTCTAAAACGGACTGACCTTCAGGTTCTATTTTATTGTTGAGTTAAAGCCTCTAAACGTTGGTCTCTCCTGCAGGAGTAAGTC
CTTCCTGTTCAGCGCCATCTACGCCACCCTGGTGTTCGGAGGCCGGCACTACATGAAGCATCGACCCAAGATGAACCTGCGGCGGCC
GCTCGTCCTCTGGTCGCTCAGCCTCGCCCTCTTCAGGTAACTCACCGCAGTGTTTCCGGTACCATTACACCAGGGGGGGCCGCCCCC
CTCCACCTGGAACACTGGAATTCCGCCCTGAAACTGAGGCAAATACGTCCGTGTATGCGCCCACAAGGTCAACGTTCACGCGTCGCG
TCAAAGTGTGCGGTCTGCGACTAAAACCTCCCTGCCTCTCCTCCCCCAGTATCATCGGAGCGATGCGGACTGGCTCCTACATGGTCC
ACGTCCTGAGCGGCAGCGGCTTCAGACAGTCCATCTGCGACCAGAGCTTCTACAACGGCCCCGTCAGCAAGTTCTGGGCCTACGCCT
TCGTCCTGAGCAAGGCGCCGGAGCTCGGTGAGTTCAGAGTGTAAGGTCATGTGATAATGTGACCTGGAGAAGGCCTCCTGAGTTGTC
TGGTTTGACCCCGTCTCTCCAGGCGACACTGTGTTCGTGGTGCTGAGGAAGCAGAAGCTGCTATTCCTGCACTGGTACCACCACATC
ACCGTGCTGCTCTACTCCTGGTACTCTTACAAGGACATGGTGGCCGGCGGCGGCTGGTTCATGACCATGAACTACGGCGTGCACGCG
CTCATGTACAGCTACTACGCGGCGCGTGCCGCCGGCCTGCGCGTGCCGCGGCCGCTGGCGGCGCTCATCACTGGCGCTCAGATCGGG
CAGATGGTCATGGGGCTGGCGGTGAGCGGGCTGGTGTACCGATGGATGCGACACGGCGACTGCCCGTCGCACCTCGACAACTGCGCC
```

```
TGGGCGGCGCTCATGTACCTCAGCTACCTGCTGCTCTTCTCCAACTTCTTCTACCAGACGTACCTGCGCCCAAAGACCACCAAGGCC
GAGTAGGGGCCGGCCTGCAGTGCATTGTGGGGATGGAATCTGCTTCTCTGAACTGAATCGGGATGATTCTCGTTTAAATGAATTCCC
GCCCACTCGGCTGTCCAACTGGTTTGAGCTGCAGCAGAAAAATATACATTTTAACTCGACGAGGCTCGGTTATCAGATTTCTTCTTT
TGGGTTCTGATCAAGCACTTTGTTGTTTTAAATGGATTCATCATCTGGTGTCAAGTGTCTTCTACTGAACTACTAAACCGATTAGTT
TACAGACGAATGAAGCTTTAATGATTAAATAAATGACCATCGATTCAACAGAAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTC
CTTTGACTTCTTGGTGGCTGAAAGAAGCAAATCTGTTTATGATCAATATTTATTGACGTCTCAAAACACTTTGTTCCACTAAATAAC
TTTATTTATGTCACATGGTGAAAATTTGTACTTTAATGTTTTGTATTTTACTAAAAGCTAAGAGACGTCTTCATCACAAACAGGTTT
AATCTTTCATTTGTTATTTTATTTGTATTGTTACTGTGTGATCGTGAAAGCTGATGATGACTTTATGCATGAACACTAGTTTTTATA
TTTCTCTTCTTCAGGGTCCAGATGTTTGTATTTATTGTTTTGACTTTACATAATAAAGAAGCACTTTTATAAATGAGTCATTTTTCT
TTCAAGCAGAAACTTGTAAATATTGAAGTTAAATTCTACACATCTGGATGCGGCTCCTGTTGGTCCAGTCGGAAGTCGTCTCTGATG
AAAGCATCTTTATCACCGAACGCTGCTGAACAAACCTTCTCCTTGGATTCCACTTTGTAGGACGAGCTGGGTGGACAAACCCGATCA
TCCATCAGCTGAGCTGCATTGGTTTAATTGATCATTATGAAAAGCCGATTCTTTTGTATGATCAATGATTTTTATTGGCACTTAAGT
ACATGTAACTCCTCACTTCATGCTCTGCTTTGCGGGGTGGTGTTCTGGACTACAGGCCGACTCGGACCTCAAAACCTGGATGACCAG
CGACTTCCTGACGCTGAACTCCGACCTGAAGTTCTCTTCATCGACTGAGTCCGCTGATAAAGTTTCTCTGGATGGAACTGTTCTGGT
ATCCGGCTGCACCAGAATCTTTCTTCATATATGTAGTACATCAAAACTCAGGACCCGCGTGTCTCAAAGTGATGCAGAGTATTTACG
TGCTTGTTCCCTCCAGACTGGGTCACTGCGTACGCCTCTCCAGCTGATTCAGAAAGCTGCAGCACGAGTATTAAATCTAATCTATAT
GCATAAACTTACTAATCATTCAGAAAACCGTGTGAAACACATTTAAATTGATTTAAGTTCAGAGGCTTCCTGTGAGGTTTGACCTCA
GCAACTAAAAGGAAGTGACATCACACAAAATGGCCCTGCTGCTTTAAGAAGAGTGAAGAAACTAACGTGACAGTTTGCTGATCTGTG
GTCGCCGCCTCACAGGAAGTCCACGCAGACGCATTCAGAGTGCATGATGGGACTGAGGACGCTGCTGCTCGTCTCCCTGTTCGCTGC
GCTGGGTAAAAGACTTCCTGTTTGCTTTGCTTTGTACGACAAACTGTTGTGTTGTGATGATTAGTTGAGAGCAGATTCGGTCTAAAG
TGGAGTGAAGGGATTTAAAGTCACGTTTTGATCCTAAAGTCAGGCTCACTTAACGCAACAGGAAGTGACTTAAAGGACCAACAAGCA
GGAAGCAGCACCCGCTTCATTAAGGAGTCAATGTTTAACAGTTCACAAATATCAAGGAGTAAAGTGAAAGGTGGTTTAATGGAAGTG
AACTTCAGCAGACGTAGCGAATGAAATCTGAGACCAGAGAACCTCGAGGATCCCGACAGGAAAGCAGATCTCCTTCGTGCCGCTGGG
TTCACTGAGGTACCTGTGCTTCCCTCACAGAGCCCATGCTGGCGGCGGGCAGCCATCGGGGCCGCTCCATGCCCGGCTCCCCGGTCA
ACATCAGCAGGGAGGACCGCGGCCTCCGGCGCGCCGTTCTCGCCGCCGCCCAGTTCTTCAACAACCAATCAAACGACGCGTTCCTCT
TCAAACCCTCAGCCGTCCTCGGAGCACAGCGGCAGGTAGGCAATTAGTACTTTAGAACGATTGTATTTTGTTGCAGTAACATTTACT
GCAGTACACGCTGATGTACTTGTACTGTGCTGGACTATAGATCATCAAGGGGATTCGGTACTTCGTAGATCTGGAGATCTCCAGAAC
CGTCTGCCGTAAACGAGACGACAACAACAGCAATCTGTCCAGCTGTGACCTTCAGCCTGCAGGTCGGCTGCAGCAGGTGAGTCTCCA
GGAGGTGTGCAGGTGCAGCTCTGAAGAGCACCGCCGACTTCCTGGGCCGCACACCTGTTTTAGACCTGAGAAGACCCACCATCCGTT
CTATGCAGAATAGTTCATATTTATATTACTAATCTCGTTTCTTTACGCCGGTTGGACTCATGGTGACACTTCACCAAAACAACAATT
AGAGGCTCTCGTGTGGAACTTCCTGTTCTTGCAGACGATCCAGTGCCACACAGAGGTCTGGGTGATCTCCTGGCAGAACGCGACCAG
GACCCAGGTGTTCCACTGCAGAGCCTGAAGCCACCTCCACCTTCATCATCAGCCTCACCCGCCATGGCCTTCATGGTGCTGGTGGTG
CCCTCCACCCGTTCCCCCCTCTTTCACACCGAACAAAGTATGGACTCGCTGACCTCAGAGCGGTGAAGGAGCGATGGGTGTCGATGA
CGCCACCTGCTGCTCAGAAGGAAAACAGCATAAACTTTACATTTATTTTTAGTTTAACTTTAAAGGGCAGTAAATCAGCATGTAAAG
TCTAAAATAATCACAGCAGTAATTCCTTTTACTTTATTAACATTAGTTTAATGTTTTTGAGGTGAAAATGAAAACCACCACAGACGC
GTTTCTGATGCAATAAAAGCTTTTTAATCAGATTTTTCACACAAGTAATTGATGTATTTAAATGAATGGTTTTCATCAAAAACACAG
ATTTGATTCAGTTTCTTTAAAGGAAGCTGAAAAGGAACCGACCCAGGTTCTACCACAGCCGCTCTGAGGAGAAGCCGCTTGGCGCTT
AGACTTCTAACCACAGCTACCAGGAAGTGAGCGGAAGTACAGATGCCATCATGCTAAATATTACTAGTGTCTACTTTACTTCATGCT
AAACGGGACCATTCATGACATAACCACGTTTTGAATCCATATCCAAGACGTGATTACGGCTCCTGACCCGTTTTATCTGAATCACAC
AAAAGAGTCTAAAAACCAAACCAACATCGACGTCAAGCAAATTTGACGGCTTCCTTCTAATTACAAATAACTTACGTTACACTTGTC
TTCTCATTTCCAAAGCTTTCATATATTCAAGTGAGAAATGGCCGCTCATCGCAGTGATAGCGCTAACAATGCTAACCCAGACGAGTC
```

```
CCTCGAGAGACGACCTGCTGTTTCTGTGTATTAATTATGAAGTCAATTGAAACATGTAAAAGTATTCACGTATGATTAAATGGTGTT
TTCTCTGCATGAAGCATTGCTGCCAATTTTATTCCGTACTGTAAAGTTGGTTGGTTTTGCTTAACCCATTTAGTCATGTAATATGCT
AATGACATTTATAAGGCTATATTGATTCAGCGGCAGCTTGCTAGCGCTGCAGTGAGCAAGCTAACCCACTATTAATTATCTTTATTA
GCTTCAGCTAGCTAGAAAGGTGTTTTTATTTTTACCTGGTTGAAAATATCTTCGTGCTGGGTCTCAGATCCTGAGCTTCAGTACAAT
ATGAAAACAACTTTGCACATACAATTTTGTCAAATTTGATTCCCAGCTGGAGAACAACTCACGACGTCCCCAAAACAGAACAGCAAA
CATGAACACAAATGCAGATTCTTAACTGCGACAACGAGGCGCTGCTGTTGGCAACCTACGGATTCATGAAGGTTGGAAAGGAATCAC
GCGTTAAACCCCTCAAAATATGCCGATTAGATCAATAACGGACCGTCGTCAAAGTCTGACAAACCACCTGTGACCTTATCACCTCCC
CCCCGAGCATCATGACATCACTGCTGAAGACAAACGGCTTCAGACGTCCCACAGAGTTGCCAAGTTCCAACAGCGTGTCGGTTTCCT
CTAGACTTTACTCAGGTCCAGTTTGGCGATGTAGGGCGAGCGGAAGGAGCCCAGGTACAGGCTCCCTTTGTGCTCGTGGACCTCGCT
GACGTAGGCCGCTACCAGGCCGTTGGGGTCGTGGAAGCTCCGAGTGCAGATGCCGCCGTCATGGAGCTCTGCCACCAGGCTGTACCG
AGGGACGAACTTCATCAGCACCTCCTGGCTGAAGAGCTGGAGAGGAGGGAAGAAAGGGATGGGTTGGGACGGCATGAAGAAGGAGCG
AGGAACCTTCTATGATCTTCTGTTTGGTTCAATATTCTCCTGTATGAAATGAAAATCCTGGTAAAAATGCATTCTACGACATTCTAG
GAAGACCTCCAGTGCGTTTTCAGGTAAGAGTCTGCAGACAGATCTCTTTCAGGGGATAAAATCCTCTTGGTCAGCAGGCCAAACAAG
ACTTTCACAGAGTTGTTTCCTGTTATGAGAATCATCAAGAACACCTACCAGTCCTTATAGTCTGTACTAGAACACCTACCAGTCCTT
ATAGTCTGTACTAGAACACCTACCGGTCCTTATAGTCTGTACTAGAACACCTACCAGTCCTTATAGTCTGTACGAGAACACCTACCA
GTCCTTATAGTCTGTACGAGAACACCTACCAGTCCTTATAGTCTGTACTAGAACACCTACCAGTCCTTATAGTCTGTACTAGAACAC
CTACCTGTCCTTATAGTCTGTACTAGAACACCTGGGAGTCTGTATTAGAATTTCTCTCTGTACCTTGAAAATGAATTTCTTGATCCA
GGGCCTCTGGGACAGGAAGTCCAGCATGGAGAAGCCGGGGTTCGGCCGGACCGACGACATGGCCACCCAGTAACCTCCGGTTGAGCT
CGGACGAATGTTGTCGGGGAAACCGGGTAGATTGTCAATAAAGGTGTCCATGCCGCCTTTGTTCAGACCAGCAACATGAACTCTGAC
GTGGAAACACAAAGACTCAGGAGCGGATCTTCTCCCACATCTCCACTCTTGGTTTCAGCCTTTTTATAAAAAGCCAGTGGCGCCTGA
CTGAAGCACATATAGAACCATGTCTCTATAAGTGGACTGGGATCAAGATCAAGGCCACACACATAATGTCACGGCTCTCTTGAGTTT
CCAGTTATTTCTACAAATCTGATTTGTCTCTGATAGAGTGATTGGAACAGATACTTTGTCACAAAAAACATTCATGAAGTTTGGTTC
TTTTATGACTTTATTATGGGTTAACAGAAAAAGTGACAATCTGCTGGGTCAAAAATATATACGTACAGCAACATGAATTAGCAATTT
TGGTGAAAGTTGTGTCAATGGAATGAGCTTCATGGCCTCTTAACGTTTTGTGAGTGATTATAAGTGACTACTGCTGGTGACTCTGAG
GCCTTTTAAATAGAGCTCATTGGACACAAACACCCACAAACGCTACAATGGGAAAGTCAAAGGAGCTCAGCGTGGATCTGACAAAGC
GAATCATTGACTTGAACAAGTCAGGAAAGTCACTTGGAGCCATTTCAAAGCAGCTGCAGGTCCCAAGAGCAACAGAGCAAACAATAG
TTTGTAAGCATAAAGTGCACGGCGCTCTTTTGTCACCATCAGGAAGAAAACACAAGCTATCACCTGCTGCTGAGAGAACATTGGTCA
GGAGGGTGAAGAGTCAACCCAGAACCACCAAAAAGATCTGCCAAGAATCAGAAACTGCTGGAACAGTTTTTGGTCAGTGTTTGGCAT
CTCCATGGACGGAGAGGCTGCCGTGCAGGAAGGAAGCTCTTGCTATAAAAGCGGCACCTTAAGGCTCGACTGAAGTTTGCTACTGAT
CACATGGACAAAGATGAGACCTTCTGGAGGAAAGTTCTGTGGTCAGACGAAACCAAAATCCAGCTGTTTGCCACAATGCCCAGCAAT
GTGTCTGGAGGAGAAAAGGTGAGGCCTTTAACCCCAAGAACACCACGTCTACCGTCAAACACGGTGGGGTAGTATTGTGCTGTGGGG
CTGTTATGCTGCCAATGGAACTGGTGCTTTACAGAGTGTGAATGTGATGATGAAGAAGGAGGATTACCTTCAAATTCTTCAAGATAA
CCTAAAATCATTAACCTGACGGTTGGGTCTTGGGCGCCGTTGGGTGTTCCAACAGGACAATGACCCCAAACACACATCAAAAGTGGT
AATAGAACGGCTAAATCAGGCTAGAATTAGGGTTTTAGAATGGCTTAAGTCCTGACTTAAATCCCATGGAGAACATGTGGACAATGC
TGAAGAAACATGTCAGAAAGCCATCAAATTTAACTGAACTGCACCAATTGTGTCAAGCGGAGTGGTCAAAGATTCAACGAGAAGCTT
GTGGACGTCTACCAAAAGCGCCTAATTGAAGTGGAAATGGCCACGGACATGTGACCAAGTACTAGCACGGCTGTATGTATATTTCT
GATCACTTTTTCTGTTAAGCCATAATAAAAAAAAATCATAAAAGAACCAAACTTCATGAATGTTTTTTGTGACAAAGTATCTGTTCA
CTCAGAGAAAAATCAGACTTGTAGAAATAACTGTGACATTATGTTCTTTACACGTGTAGGTAAACTTTTCACCACAACTGTATAGGA
CCATGTCTCTATAAGTGGACTGGGATCAACCGCTGCAGTAGTCCATAATAAAAGGGGTCAGGTGGGCTAACCTGCGTATCCGGGCCA
TGGTGGTCTCCGCCACCAGCACAGATTCCTCATCAGGGAGCAGCTGGATGCCGTTTGGGAAGCGAAGATTCTCCATCAGCACCGTCA
GTTCTCTGCTCTCTGTGTCGTACTCCAGCACCCTGACGGGAGACACGAGGGGCTCTTATTGCTAAACAACAACAGGAAGTCAGACCT
```

```
TCTTTACCGAGGAGGTTTAGAACGTAAAAAGACCCAGAATCTAGAGACAAGTTAAATACAACACTGGAGGTAATTCATCCACTGACC
ATTATTTTGACTATCAGGTCAATAATCTTTAATCTCCCCGTCTGAATAATCTCTGATCTCTTCTTCTGAATAATCACGTAATATGAA
TAATCTCTGATCTCTTCTTCTGAATAATCACTTAATATGAATAATCTCTGATCTCTTCTTCTGAATAATCACTTAATATGAATAATC
TCTAATCCCTTCTTCTGAATAATCACTTAATATGAATAATCTCTGATCTCTTCTTCTGAATAATCACGTAATCTGAATAATCTCTGA
TCTCTTCTTCTGAATAATCACGTAATCTGAATAATCTCTAATCCCTTCTTCTGAATAATAACTTAATCTGAATAATCTCTGATCTCT
TCTTCTGAATAATCACGTAATATGAATAATCTCTAATCCCTTCTTCTGAATAATCACGTAATCTGAATAATCTCTAATCCCTTCTTC
TGAATAATCACTTAATATGAATAACCTCTGATCTCTTCTTCTGAATAATCACTTAATATGAATAATCTCTGATCTCTTCTTCTGAAT
AATCACGTAATCTGAATAATCTCTAATCCCTTCTTCTGAATAATCACGTAATCTGAATAATCTCTAATCCCTTCTTCTGAATAATAA
CTTAATCTGAATAATCTCTGATCTCTTCTTCTGAATAATCACGTAATATGAATAATCTCTAATCCCTTCTTCTGAATAATCACGTAA
TCTGAATAATCTCTAATCCCTTCTTCTGAATAATCACTTAATCTGAATAATCTCTGATCTCTTAATCTAAACAATCTTAATCTGAAT
AATCTCTTCTACCGTCCGTCGGCCGTGGCCTCCATGATGAGGTGCATGTAGTCTCTGCGCTGCCATCTGCTGCTGGAGTCGGTGAAG
TAAATCTTCTTCCCGTCTTGAGTGACCGCCACGTCGTTGATGAACGACAGCTTCCTGCCGGCGACCACCTGACCCCCTGACACCAAC
TGGGTTGCCTCACCTGACACAAAGGAACACAGCAGAGTCCCAGTATCCATGGTTACATACAAATGGTTGTAATGTGGTTATTATGGG
GTGTTCATGGAGAAGAGAGCTTCAGGTGTTGATGTGATGGTCTCATTACATCGGTTTCTCTTCGGTGAACGATGAGTCGCATGATGC
TTCATGTCTGTCATCAGTTGTATTTATTCCACTTCATTTTATTAATATAACAACTTTTACACCTCAGACGAGAGGAAAACACTTTTC
ATTAGGAGTTATCGAGTTTTGACATTTTGACACTATTTAAATTCATCAATAAACCCGGTGAGTTTTCTGCTTCTTCCTTTCCAGAGA
TTATTATTATTTATCAAAACTAAAACTAAGCACAAATGTTTTACTTAAAAAATTAAAAAACAAATGTTCTACCAAACATCAGCGGTC
TGAGAGAAACTGATTCTTCCCAATCACGCTTCTTCTTCTTCTTCTTGCCTGAAATATAACTGATTCCTGATACCTGTGATGCTGAGT
GGATTTAAATGTGCCTAAATGTGAGAAGAAGAATCAACCAGCTAACAGGGATCCACAGACAGACTGAGACGTGTCACAAGGGACATA
ATCAAAGCACAGAGGAGAAGCAGAAGGGGATTTCATTTTTTCTTATTGATCCAGGACAGATCACTATCATCAAGCCAAAATAAATAA
CCAAATAAAACAGCTCATCCAACCACAGAAACAACACATCAAACATGAAGGATCAGCTCTCCCTCCTCCCCCACGGCGCATAACCCC
CCCAACTAAAAGACAAACCAAACAAGTCCACTGAACAAACAGATGAATAAAACAGCCTCCCCGGCTCGGGATCAAGGTGCACTCTGT
ACCTACTAGTGGCTTCTGCCCTGTGCACCCGCTGCTCAATGCGCCGTTCATACAACAACCGCCAGCCGCCTTTCCGACCCAAATAAC
CCGCCAGCCTCAAGTCCTTCATCCCTGCCAAAGAGCGCAAGCACAAACAAACGCAGGTGTGGTGGACAGCTTTTTTTCCCTTGTCCT
GGAACTTCTCCAGGTGCGGGGTTGTGAAGGTGAGCAGCTGACTTCCCTCCTGCCCCTCCCCCAAAGACTGGCTGACGGACAAAGAGG
GGAAGCAGTGCTCACAGCCGTCACACCACTGGTTGCACAAAGTTCAGTTTCCCGCAAACTCTGTGTGGGTTCTGGACAGGCCGGCAG
AGACTCTATAGATCCGCCTGACTGAGGTGAGGCTGCTGCGGTCTCTCCGGCCCGTCATCAGGTGGCCAGTCTCTCTGAGTCGCTCAC
CGGGCTGGCGAAGTGCACCTTCTGACTGCTGACAAAGACACAAACGCCATCTGCACACGCAGATACAATAGGAGGACGATATAAACG
GACGGACCCTGGCAGTGAGAGACGTGTCTCTCAAGGGGAAAATGTACATAATATGGTTGGTTTCAATGTACCACTTTGCCCAGGTTG
CCCTTGAGTGTTTTGGACTTTATTATTATGGCTTCTTGTAGTCAGTGCGGAGGCTTCCAGCAGGGAAGAACCGCTCACTGACAGGAG
ACCAGAAGACATCCTCTTAGAGTTTCCATGTCCTCTCCACATAGGGACCCTGTCTCCACATGACAGCTGACTGTAGGACAAACTAGC
TGCTGCTTCCACGCCTTCTTCTTCTGTGGTCGTTCAGTAAGAGCGCTGCTGTTCTCTTTAGCATCAACCTGTTGTTCCTCCACTTGG
CGCTGCTCTTTCTCTCCAATGTCTTCATCAGCAGCAGCCACGCCTGCAGTACCGCAGTTCGCCAGTCGTAGTTTGGACCACAGGTTT
GTCATCCAACGGATCAACAAAAGGTCTGCTAGCCGTTGTTTCGACCACAGGTGCATCATCCGTCGGTTCAACCACAGGTCCGCTACC
TGTAGTTTGGACCACCGATGCTTCATTGCCGGGCATCGACGGCTGCTTACTATGCGGACAGGAGAACCGCTCGTGTCTCAGACTACA
TGGTGGAGCTAGCATCCATGTCCCCTGCGTTTTGGGAGCTAGCGTCCACCAGGTAGGCACTGTCTCCGTACTTTACTCGGAAGGACA
CGGGGTCTGTCAAGGAGTCCAGGAACACGAACAGCTGCCACCTCGGAGACTGGACGCCACAATTTGGGATCCTTATACCCCAAAGCC
ACTGTTTTTAACTCGCGCTTCCATCATGGAAGAAAATCCCAGCTTATTGTGTCCGGCCTGCTCACCGACAGCCAGAGACACCTCCTC
CCCCGTCCTGCTGTCACCGGGGGGGGGACTTCCTGCCGGAGGGACGGAAGGGGGGGTCGTCTTTGAGGACGCCATTCCTCCCGAAAA
ACACCTTGACTAAGACCGGATTTCTCCACACACAGAGCAGTGAAACAGTACTCAAGACTCACCGGATAATGAACATCAGCAAAGGAA
GGAAACACAGTTAGAGACCCAAAGATCTGCAGCAGTCAAATCAGAACCGAAAAGAGAAGGAGAAGAAGAAGCTGAAGAACAGCAAGA
```

```
AGGAGAAGTAGAAGGAGAATGTCCTGCCAAGGCAGCAGCTCTCTTACCTGTGGTGGGGTTCACCTGGAACAGCCCCAGGTAGGCGTC
GGCCACAAACAGAGTCCCATTGGGTCCGACTCGGATCCCCAAAGGCCTCCCACAGCTGGACTCCTCCTCTCTGGAGCCTGCAGGAAA
CAGTCAAAGCTGACCTCCACCTCCTTCCTCTCCGCACAGAACCTTCAAACATTCATGCTACTTAAGATTGAGATGATCAACGAGAAG
GAGGACCTTCTGTGGAGCAGAGTGGTTCAGAGGAGCAGCAGGAGGCTCACAAGACCACCTGGTTTGTCCTGGTTCTGTTGGCTGGAA
CCGCATGAGTCACAGAGAGATGTGTCATCTGGGAGTTTACGGTTTGAAGACCTGCCACAAGTTCCTAAGTTCTCTCTAAAATAAAAT
CTTCTTCCTTGAACTCACTCATTGCTTTACAACAAACAATCTCTTGATGTTTGTGATCCTCTTTAAGCGTTATTTTATTTCAGATAG
CAGTAGTTGTGTTTTTGTATCGCTGACGTCCCGTGTTTGATGCTTGAACTCGTTTGTGTGTTCGTGTTGAAATCAGGAAGCTCTACA
GGAAGAAGAAGTCAGCGCAGGTCTCAACACGTCGCTGTGATCGTTTTACCAACTGAATGAACTAAACTCAGGTCAACAAAAAGGGAT
TCAACTACACATGAAACCAAACTACAAATAAAAAACTAAACTAAACTACAGATGAAACTAAACTTTAGTAAACTTCCAAGTGAAATG
ATACTTGCAGTACTTTTTGGTTCATAAACGTTAAATTAAAGAAGAAGCCCTTTTTATTGCATATTGTATCGTACGTACAGGGGTTTG
TTCCCTGGGTTAAACACACACAGTAGCAGCAGTGGGCTGCCAGGGGACAGCTGGGGGTCGGGTCTCTTGCTCAGTTACACTTGCCGG
TGTTGTAACCGCCAACCCTGCGGTTACCGGCGCACCCTCTCTACCGCCACCTTTAGCACTGTTAGCACCGCAAGCTGCAAGCCGCTA
GCTCGCCGTCGCCATGTTGAGAGCCATGTGGCGGCAATTGAAATGCTCCCATTCCTGCAGCTGTATTTAAAATGAACCGTAAACGGA
TCAAATGTTTCTCTCACCACAGGGAAGCTTTCCGAGTCTCGTCACCGTGTGAATCCTTCGACCAATCAGCTTCACAATCTTCCCATC
AGCTGTTCCGGCATAAAAAACATCTGCAGACAGAGGAGAAGAACATCTATCAGCTTCTAGTCTTCTTCACGTCTTTATCATCATCCA
GTTTTACACTAAATGAAACGTTTAGCCTGGACGAGTGTGGGGCCTCAAAGGGGAGAAAAGCTGAGTGACCTCTGACCGTCAGCATGT
GGAACTCCCCAGTTTGCTTCCCCACATGTCGGACGGCAGAAGAGTCAGAGGTCACTGTTTCTTCTCCAGTCACAATCCACCTGAACT
AGAGTTCAACTAAACTAGAGTTAAATATACAATGAGCTATGAGCTATTAAAAACCAAACCAAGCCAGACCAGACAAAAGCAGGCTGC
TGCAGACCTCCTATGTTGACGATGGACTCCGGTCCGGTGATCTGGTCCTCAAACAGCCTCTCAGCCTCTCGGAGCTTCAGGTTTGGT
TCCCAGCAGCCCTTCATCAGCGGAGGCTCCTTCAGGCTGCAGTACACAAACACACCAGCAGGAGACATGCCGAGGGTCGGGGTATTG
GTACTATAAAAACAGGACTAATACTACCGTAACAAAGTCAATTTATTCTGGAAGTACTACACTTTGTCCCCAGGTGAACCGCATAGA
GACTGCAGCCTGTGGCTGCCTGTTCTCACATCCTCAGAACATCAGCTTTATTGTGAAACGCTCTCACTTGCTCCATTTAAACGTCTT
AGGTCCTCTCAAACCTTCAGAGCCTCCAGGTGTTTCCAGAAGCTCTGGCATGGGACCTGTCCTGCTGTGCAGTAACTCACCTGAAGA
CCTCTGGGTGGATGGGTGACTCTATGATGAGGATGATAACCAGGAGAGGAAGAAGCAGGAATCCGCCCAGAGAGAGCAGAGTCACCT
GGAAAACCTTCCCGCTGTAGGTGCTGAGAGAGGAACCATCAGGTTAAATCCTTAAACAAACTCAATTAACAAACCAATCAATGCTGA
ACTAGTCATCAATACTAATCCCTAGCACACACAAGGCTAGCCAATGCTATGCTAATCACTAGGATTCACAATCTTCTTTTATCTACG
TACTAAATGAAAATCACGGTGTATTGAAGAAGACGTGAAACTAGAGACTGAGACCATAAACTCATGTTTACTGAGGGAATAAATCAA
GAGAGAGTCATTTTCCCATAGACGTCTATGGGAGCAGAGGAGTCGCCCCCTGCTGGTCACTACAGAGAAGTAGAGTCATTTCCTCAT
AGACGTCTATGGGAGCAGAGGAGTCGCCCCCTGCTGGTCACTACAGAGAAGTAGAGTCATTTCCTCATAGACGTCTATGGGAGCAGA
GGAGTCGCCCCCTGCTGGTCACTACAGAGAAGTAGAGTCATTTCCTCATAGACGTCTATGGGAGCAGAGGAGTCGCCCCCTGCTGGT
CACTACAGAGAAGTAGAGTCATTTCCTCATAGACGTCTATGGGAGCAGAGGAGTCGCCCCCTGCTGGTCACTACAGAGAAGTAGAGT
CATTTCCTCATAGACGTCTATGGGAGCAGAGGAGTCGCCCCCTGCTGGTCACTACAGAGAAGTAGAGTCATTTCCTCATAGACGTCT
ATGGGAGCAGAGGAGTCGCCCCCTGCTGGTCACTACAGAGAAGTAGAGTCATTTCCTCATAGACGTCTATGGGAGCAGAGGAGTCGC
CCCCTGCTGGTCAGAACCAGATGTTCTTTCCTGATTGGCGCTGGAGCTCATTTGAATGGGTTCATATGAGAAACATTAACACACACT
ATGAGTTGATAAACAATCCACATCTGTATCTAAAGCTGTCCGATAAATGTAAAGCAGTAAAAACAACAATACTTCCTTCTTAAACGT
CGACGAGTAGAGGTATTGTAGTAAAGTACCTCTTATTTAGTATAGTGATAGTAAAGGTACATCAGTATAATTTCTCCAGCATGAAGC
TCGATGACGTAAATATCTAATAACATAATAACGATCACCGATCGATGCCTCGGCGGCTCACGAAGGTGACGTCACGGCGACTTTCGC
CTCGGTGTTTCGGTTTTAGAGTCCACCCCGGGATGCGGATCACGGAGTGGTCCCGGTTCCTTTTCTTACCCAGATCCCTTGTTCCGG
TGCTCGGGCAGCTCGTCGGTGATGACCTGCGGTCGCTGCACCCTTCGGAGGCGCAGCCCCTCCGTCTCGTTCATGCTGCACTAAGTA
CTCGGACGGTTGACCCGGTTCGGTTCGGCTCGGGTCAGCTGTTGGCCCGGGGTGGCTCAATGTAGTGAGGTTTAACCCGGGTACACA
CGGACTAAACACGCAGAGGACCGAACGACTGTCATACCGAAGCTGTTGACGCGGACGGGTGAGAGCTTCCGGTTTGGGACTTCAAAA
```

```
TAAAAGAAAACACGTCTTCGTTTTTTCTCTGTTGAAAGTTTTGATGAAAAAAATGAATAAATTCCTAATTAATCCAATTTAAATAAA
AAAAATTAAGACAGTGTAAAAAGGCAAAACTTTCACTCGTCACTTTCTGTTCGCTGGTGCACTTTCTTTATAAAATATTTGTAACTT
TTAGCTTTTATTCCTTTATCTTTAAACGAACCCATAGCCTTATATTCTATTTTATTCCTCCTGCACTGTTGTCTTGTCGTCCATTGT
CGTACGACCTGCTGTTACCACCAAGCGCCGAGTCAAATTCCTTGTATGTCTGACGTATTATGGAAACACACCTTTCCTGATTCCTGA
ACAACTACAAACTGATGTTAAAGAACTGAAGTTTTTGCTGAACTTTTAAAAGGGGGAGACATTTAATGTCACAGCTTAAAATGTGAA
ACACCAAAAAACACGTTAAAGATACCTGCAACCAATAGTTTTCTAAAAAATGAAGTTTAAAAAAGTACCATAAATATTTAAATAGAG
CACACATACAAACGTTGAAAGATGACAATCAAGTGAATTCATTTCATTTTATTTTGAACCTTCAGCCCCCCATCCGACATGTGTGGT
ATTCATTTATGAAAGAACATCAGCCTGAATCAGAGCAGACGACTCCTCTGCTCCCATAGACGTCTATGAGGAAATGACTCTACTTCT
CTGTAGTGACCAGCAGGGGGCGACTCCTCTGCTCCCATAGACGTCTATGAGGAAATGACTCTACTTCTCTGTAGTGACCAGCAGGGG
GCGACTCCTCTGCTCCCATAGACGTCTATGAGGAAATGACTCTACTTCTCTGTAGTGACCAGCAGGGGGCGACTCCTCTGCTCCCAT
AGACGTCTATGAGGAAATGACTCTACTTCTCTGTAGTGACCAGCAGGGGGCGACTCCTCTGCTCCCATAGACGTCTATGAGGAAATG
ACTCTACTTCTCTGTAGTGACCAGCAGGGGGCGACTCCTCTGCTCCCATAGACGTCTATGAGGAAATGACTCTACTTCTCTGTAGTG
ACCAGCAGGGGGCGACTCCTCTGCTCCCATAGACGTCTATGAGGAAATGACTCTACTTCTCTCTTGATTTATTCCCTGAGTAAACAT
TGTAAACATGAGTTTACGGTCTCAGTCTCTAGTTTCAAGTTGTCTTCAATGCAGCATGATGTTCATTTAGTAGATTATGGTCCATTT
AGAGTCAAACAGACCATAAAGAGGGGATGCTTTTGGGCGGGACTACACAGCGACTGACAGGTTTCCAGAGACTAAAGCATCTCGATG
TCCTCCTCAGTCCAGATGAGGTCAGACTACCCCACCTCTAATAGACCACATCATGTAAGGTCCTCAGAGGTCCCCCCCCCCCTCGGC
AGAGGCACACTTGTGCTCTTTGAGGCTCCTGGAGGAGGCGAAGCGCCGGCTGCAGGAGCAGCAGCTGAAGGGTTTAAGTCCCGAGTG
AACCAGCAGGTGAGACTTCAGGTTGTTGTTCTGACTGAATCTCTTCCCGCAGACGCCGCACTCGTAGGGTTTCTCTCCGCTGTGGAC
GCGTGCGTGTCTGACCAGGTTCTGTGACGCGCTGAAGCTCCGCCCACAGACCTCGCAGACGTACGGCTTCTCTCCGCTGTGCGTGCG
CTCGTGCAGCTTGGCGCTGGCCCGTGCCGAGAAGGTCTTCGCGCACCGGGCGCAGCCGAAGGGCTTCCCCTCCGAGTGGCTGCGGCG
GTGCGTCCGCAGGTAGCCCAGCGAACTGTAGCTGCGGCCACACACGTCGCAGGAGAACGCAGCGCCGCCGGCGTGCAGCCGCCGGTG
GGTCTTCAGGATCTGGGCGCTGGTGAAGCTCTTCCCGCACTGGTCGCAAGTGTGAACCTTTCGCCCGCCGTGCGCCAGCTGGTGGCG
TCTCAGCGACGCGGCCGAGGAGAAGCTCTTCCCGCACTCGCCGCAGCCGTGCGGTTTGATGCCGGCGTGGCCGCGCTCGTGCGCCAG
CAGCGCGCTGCGCTGCCCGAAGCCCTTACCGCAGGCCTCGCAGATGAACGCCCACTCGCCGCTGTGCGCGGCCCGGTGCGCCTTCAG
GTGCGCCGCCTGCGTGAAGCCACGGCCGCACACGTCGCAGCTGAAGTGGCGGCGGCCCGAGTGGACGAGGCGGTGGCGCCGCAGATT
GCCCGCTACGCTGAAGCTCTTCCCACAAGTGTTGCAGTCGAAGGGCCCTTCTCCCCCGTGTGGACGCGGCGGTGCACCGACAGGCCG
CTGCTGCTGGAGAAACTGTTTCCACACTGCTCGCAGGTGTGTGGCCGCTCCCCCGTGTGCGTCCTCATGTGCGCGGCGAGGCCTCCC
CGCTGGCCGAAGGTCTTCCTGCAGACGCCGCAGGTGAACGGCCTTAGTCCCTTGTGGACCGTCAGCTGGTGGACGTGGACGCTGCGG
CGCGTGGTGAACCCTTTCTTGCAGGTTGGACATCTGAAGGTTTCTCCGTCAGGTGGACCCGCCGGTGGGCGGACAGGTAGCTGCGCT
TTGAGAAGTCCTTCCCGCAATCAGCACACCTGAACCCGCCGGCCGTCACCCGGCAGGTGAGCCTGTGCGCACGCTGGCAGCACCACG
CCGCCAGCGTCTCACCACAGGCAGTGCAGGCGAGCGGCCGGCCAGGTGCCAGGTGCGCCCTGAGGTGGCTGCGGGTGCCGAATGGCT
TCCCGCAGTGGGAGCACCTGAAGGGACGACAATCGCCGTGCACCGTCAGCTGGTGCACTCGGACGCTGCTGCGATCGGAGAAGCACT
TCCCGCAGGTGGAGCAGGTGAAGGGCCGCTCTCGGCTGTGGACCCGCAGGTGGGGGCTCAAGTTGGTCCGTCTGGAGAAGCTGGTTT
CACCTGTCTGAGCATCTGATTGGTGGATCATCTCCAACTGGGCGTCTCCTTTAGCTGCAGTGGAGGTGGAGTCATGGAGTCTCCTCT
TATCACACGATGCTAGTTTCTGCTCCTGGTTCTCCTCCAGCAGCCTCCTCAGATCCTGCAGCAGAGCCTCACAGCACCTCCTCATCA
CAGCCAGCAGCTCCTCCTGCAACACCAACGGCTCAGAGTGTGAATTACTATGTGTACATCTGAGATGTACTGCAATATCAGGTACTA
TGCATACCTCTGTGATGTACAGTATGATGTACTATGTGTACATCTGAGATGTACTGCAATACCAGGTACTATGCATACCTCTGAGAT
GTAGTACAGTATGAAGTAGAATAGGGAGGAGAAGTATGATATCCTCTAGTACAAGTACCTCAGACGTGTACTAAATTACAGTAGAGT
ACTTGCACATGAATCAGACTGACCTCCGTCCTGGAGTCTTCTTCTTCTTCTCTGTCTTTAAGTCGCTGACAGACTTCCGTCACTGAC
GTCACCATGAAGCGCTTCACAAGCTGCTGCACAGAGACCCACACCTGATCCATCACTCATCCACCCGTCACTCATCCACCCGTCACT
CATCCACCTGTTCAGCACGTAGTCCGATCCTCGTCCACCGGGCGGCCATTGCCGCGCCTCCACTTACGGCTTTCTTTACGACAGCAG
```

CAGTGAAGCCCCCTGCCGGCCGGAGGGCGAAGTACTATAACAAAGTCGTGTAATATCTACAACATTATACATGTTAAAATAACTTGC

AAGACTTTGCAAAAATAACGGAACACTGTGAGAAATGATATCAATAAAAATAATATTATTATACATCTTTATTTATTCACATTGATG

ATTCACAATGTTTCAAAAATAATTTCAGCTGGAAACTATAACTCTATATAAAATATGTATATGTTCACAAACATGAATTAGTTAAAT

AAAACATCTGGAGGAAAACAGGAGACAGAAGAACTTCAGAAGAAAGAAAAGCTTCAACTCTGTCTGTTGTTATCTTTTATTATGAAA

TACATCTGTATCAATCTATTTCAAAATGAGCAGAAACACTGTATTTCATTTCTAATCTTGTATTTTTTAAATGTTAAAATTCAGATT

TTCATTTAGAAAAGAGGACTTTAGTCTAAAAACTAAAACACATAATTTCTTTCACTTTAAAAGGTTTACGGTGGATTTGACATGTTA

CAAACAGAAGACACAAGATTGTGTGTGTGCACAGCGGGTCAACCTGCAGAGCGATGATGTCGTACCTGAGGACCTGCTTGGTCCCAT

CCAATGACCCCCCCCCTCCCCCCCCTGTGTCTTTTAAACAGGCAGCAACGGGAAGTTCTTCTT**CTTATCTCCCCCCCTCT**

**GCGTTCTATACTGCTTGTGAGCCTCGATGATCTCCATGACAACAGACGGGTCGTCCAGGGTCG**

Exon 15 **ACACGTCACCGAGGTCACCTGTGGTCTCCATGGCAACCTTCCTGAGGATCCGCCTCATGATCT**

**TCCCGGAGCGAGTCTTCGGCAGCCTCTTCACAAC**

CTGAGGGACAACAGGAGATACTTTCAAGTGTAAGGAGCCGCTGGAACAGCTGGAACACCTGTGTTTCCAGGTGTGCATCCAGGTGTT

CCTCAC**CAGGAAGTGCTCAGGAACGGCGTATCTGGCGATCTTTGTGGAGACGAGATCTCTGAGCT**

Exon 14 **GCTGCAGGACGAGGGGGGGTCAACTCCAAAGTCCTCCTTCAAAACCACGAAGGCAAACGGTA**

**CT**GAGGACACACACACACACATAGTGAAACACACACAGATAAACACACAGAAAAAGAAACACACAAACACACACGATAAAACACTA

CGGTGGTTGAATAAATTATATATAAATGAATTGATTACTCAGAGATGTTAGATGAATGAATCCCTGGGCTTTGATTTGCTCCTTATT

ATTAATATAGGACATACTTTTCTTTTTTACTGTACTATACATATACTGTATGTACTATATATGCGTATATTTAGGTAAGATACCAGC

AGGGGGCAGCAGAGTCTCATAGTTCAGCTTTTATTTCTCGTTTTGATGAAATACAACGTGTCGCTCTGACTGAATTAGAACATTATT

TGATGATATATATATATATATATATATATATATATATATATATATATATATATATATCCTCTTCGAATAGGATCTCCGTTCACGT

CTTCATACGTTTTATCTCATTCTGTTTGGAAGTGTAAATATTCTATGCTTGGATGTTTTAATATTCAGGGGGGTGTTTTCTAAAAAT

TTCTTAATAGCTTATTTTTTAATGTCCACTTTTTTAAAATATTTATTAAAGTTTTAATTTAAAAGCAAGGGGGGGGGGGGGGGTGG

Exon 13 AGAAGCAGAGGATCATGGGAGTCACCTCTCACC***TTCTCCTTTGATCTCATGAGGGAAACCGATCACAGCTGC***

***TTCTGGAACAGCTGGATGCTCGTC***CTGCAGAACAACACCACGCAAACGTGTCAAATCCGATACATATATGTTAATCAT

ATCATATAACTCTAATTTCAATCAATATTTAATTCTACATTTATTTCACTTTATCTCATCGTCTTATTTTTGATCTTTTAATTGAAA

CATTAGAGTGTAAAATGTCCCGTCAATCATTCCTTTATTTGTTTTGAGCTCGTGATGAACGCTGTTTTTCTCTATTTGTTTG

**cDNA sequenced reverse**

TCTGGCGATCTTTGTGGAGACGAGATCTCTGAGCTGCTGCAGGACGAGGAGGGGGTCAACTCC

AAGGTCCTCCTTCAAAACCACGAAGGCAAACGGTACTTCTCCTTTGATCTCATGAGGGAAACC Exon 13

GATCACAGCTGCTTCTGGAACAGCTGGATGCTCGTCCAGCGCGTCCTCGATCTCCGCGGTGCC

Exon 12 GAGCCGGTGGCCGCTGACGTTGATGACGTCATCCATCCGCCCGGTTATCTGGTAGAAACCGTC

CGCCGAGCGCAGCGCTCCGTCACCGGTGAAATAATAACCTGGAAACGGCTTGAAGTACGCCTC Exon 11

CAGGAAGCGCTGCTGGTCTCCATGGATACCGCGAGCCATGCCGGGCCACGACTGGCTGATGCA

CAGGGCCCCGCCCACTCCGTCACCCAACAGAACCTCCCCCTTCTCTCCCAGGAGACTCGGCTG

Exon 10 GATGCCAAAGAACGGCCTCATGGCCATCGCAGGAACAATGGCGGCTCCTTCCTCTGCGGGTCT

CGGGGCGATGCAAACCCCCCCGGTCTCTGTCTGCCACCAGGTGTCCACTACAGGACACCGTCC Exon 9

Exon 8 CTCTCCAACCACACTGTGGAACCAGTGCCAGGCCTCGTGGTTAATTGGCTCCCCGACTGATCC

CAGCGTCCTCAGAGACGAGCGGTCGTACTTCTTCACAAAGCTCTCATCGTATTTTAGCAGGAG

ACGAAGAGC

>groupVI dna:group group:BROADS1:groupVI:16660651:16676786:1

CATGAAGTCCTCCTCCGGGTCCAGGGTACC***TGATCCCAGCGTCCTCAGAGACGAGCGGTCGTACTTCTTCA***

**Exon 7**

***CAAAGCTCTCATCGTATTTTAGCAGGAGACGAAGAGCCGTCGGAGCTCCGTAGAACTGAGAGA***

***TCTTCAGGCGCTGGACGGTCTCCCAGTACCGA***CCTGGAGGACACGGAGGGGCGGACAGATTTATTGACCTTTATG

TCTTTAAAAAAACAAACCGGTACCACACGAATACAACAAAACAAACAATCAAGAACAGCGTCTCTTCACAGTGTCCTCATACACCCT

GTTTGTCTTCAATGTTGCACCGTTCCTCAGTGAAAACCTGCATGTGGTCCACTGGGATGTTCAGTCCACAAAGGACGACCTTCAGGA

CCCCGTTCCTCGTACGTGGGTCAACTCAAGCCTAGTTTATGCTTCTGCGTTTTCAGAGAGACGCAAGGACACGCAGACGCAAGACCC

CCCTGCGTGTCCCTCGCGCGTCTTTTCACACCTCCTTGCGTGCGTCGACCCATTTTTCTAGACTAGCGTCTAAAGCGGCGCAGCCTC

CTGCAGCACCAGGCTGTGATTGGTCCACTGACTACGTCCTTTACGGAGTCTCACGTCTCCGTTTTCACAGCACAATACGGCCGTTAT

TAAAAGGAGGAACGTTTACGAGAGAACGGATCAGATTGAAGAGCTTTTGTGGGAAGAAATGCGTAAATATGAGCGCTTGTACAAGCC

GTCATTGGCGGACTATGAGGACGCCGGATGGCCTCTGATTCATGGAGGGAGATCTCCACAAACGAGGGGAACAAAGTCGTGGAGGAA

AATACGGGACAAATGTGTCCTTGATGAAAAGCAGCAGTGTGGATGCACGGGGCAATAAAGTCTATGATCAATAATTAAAGCAACCAG

CGACTTCCACACACAAAGACGTGACTCTCCAGGTCGTCTTCAACAAAACACGTGACTCCACCTGTTGTTCTGGAGGTGAATCGCTCT

GCAACACACGCAGGACTTTACAACCAGGAAGTGAAACCAGTGCGTCGACGCGTAAAGACGCAGAAACACGCGGCGGCCGTCAGTTGA

TCAAATGTCCCATTATCCAGCTCAAATTATGATGCTTGACATCAGACTATCTCAGTTCCTCGAAGGCGGATCCGCGCTCAAACAGTC

TCGTTAGTTTGAAGCAGTCAGTCGTGTGCGTCCTACGTCCGAGGGCGGACGAGTCGGTCACCCAAACCAGGATATTTTACAGCAACA

ACGGCAAATAAACGATGTCACGGAGAACTCTTTCTCTGCAAAACGTTAATCTAACTCACATCGCTTCTTCATCAATGAGGAAGAGCG

CTGCCTTTTTGGGGGGCGAGTGTTGAACTAATCTAAATTGTAATGGTTCTTATCTGGGGGGGGGGGGGGTGTAAAGGTGATTTAAG

TTGTAAATCTGCTTATTTGAGGGAATTGCACTTTCTTGTTTCTTGTTCTTCTGAGTTTGTGTCCTTATGGGTGAACTGTTTTTATTG

TCGCTTTGGATAAAAGCGTCAGAAAAATGACATGTGATGTAACATCTGACCCATCATTATTTTGTAGTTACTTCCTTTCAAACCTCT

GTGGATCAACCGACTACCAGTCAACGGTCAGAACCACCGCAGGGACTAAAGGTCACATATGGTCTGAGGAGATAAGGAGGCAACAGT

TACTTCATTAATCAGTAACCAGGAGACATCTCATCTCAAAGGTTCCTCCGTGGTCATAAAGCAGATAACAAGTAGAAGGATCCGTCA

CTAGATAGATGGATCAATGATTGATCAATCGACATATCAAATGTCTATCGTCCTCATCTATCGATTATAGTTCTAACTATAGTTTAA

TCTGTAGTTTAATCTATGGATTGATCTATTGTTTAATATGTCGTTTGATCCATCGTTTAGTCTGTTGTTTGATTTGTCTTTGATAGA

TCTTTGATCGATCATCGTCATGAACACCAGTTTGGACCCTCCTGTGCACATCAAAGTCTTCTCACCTGGACCTTA***CCTGGGTC***

**Exon 6**

***GGGGTAAACAGGTGTGCTCTCGAACAGGACAGTGGTGGATCCGTTGCATAGCGGGCCGTAGAC***

***CACGTAGCTGTGCCCCGTGATCCAGCCCACGTCCGCCACACAGCCGAACACGTCGCCGTCCTG***

***GTGGTCGAACACGTA***CTGAGGAGACATCACACAAGGGGTCACTTCTCTCTGGTTCTGTTGAACTACATTACCCAGAAGCCC

CGGACTCATTTCATTCATTGAAGTTCAGGACTCTTGGTCTTGCTGGGTATTTTGAGGAAATAAAGATAAATAAAATCTCTCTCTCTC

TCTCTCTCTCCCTCTCGCTCGCCCTGGCGTGGAACCAGCTGTCCTTGTCTTTTTCAAATAAAGCAGCTTTTGTTCCCTTGATTCATC

CAAACAAACAACGAGCCAGAAAGACAAACCCTGAGCTGCTAAGCTGCTTTAGAAATTAGCACCCGAAAATATTTTGGCAGTTATGTT

GATGGTGGGGGGGGGGGGTTATTATGCTTTATGCTTTTATTCAATAATCATCATACAAATCACGCGAATCCCTGCAAACACACACAC

ACACACACACACACACACTAATTTGTGCTCCTCTTTGTCTTCCTGTAATCAATGTTCCATGTTAATACTGCCACAATAAGAGCGGAG

AGAGGTGGTGCAGGAAGGTGGGACAGGAAAGGCAGGTGGAGGTGGGCAC***CTGGTGAGTGAGTGAGGTGTACAGCAG***

**Exon 5**

***GTAACCGGCCTGCGTGTGGACGATGCCTTTAGGTTTCCCCGTACTTCCTGATGTGTATAGAAG***

***AAACAAAAGATCTTCACTGTCCAAAGGCTCCGCGGACAGACGGGAGACTGAGAGGACATCGC***

CTGTAAGACAGGACAGGTGACAGTACACTTACCGCTACTACTACCACCGTCACCACCGTCACCACCGTCACCACCGTCACTTTAGTC

TATTCGAACACGTCTGCTCTGATATTTTATTGTACTGTGTATATATTGTGCATTTTTTATTTCTCATTGTAATTACTCTACAAATTC

CAGATTAATGAAGTATTTATCTATCCACTACTATAACTACTAATACTACCACTTCTGATTATACTACTACCACTACAGGACAGACAG

GAGACTGAGAGGACAATACCTTTACATCACATCACAGGTCATTTATCTGACGCTTTTATGCAAAGCGACTTACAATAAGTACATTTC

AACCATTCGGTTGTCCTTATGGTCCTTTAAGTTTCTAGCTGACAATCAGGTGTATTGATCAGGTGTCTAAC***CTCCTCCAGGG***

**Exon 4**

***GGACGTCCAGCTGTCCCATCACAGCAGGATTTTCTGTCCTCTGAGCTACAAACACATGTTGGA***

***CGGTTGGACAGTTCTTCACCGCTGAGTCCACCGTAGCCTTCAGGTCAATGAGCCGTCCTCCCC***

***TCACCCCTTGATTAAAGGTGACCACGGCTTTACACTGGG***CTGTGAACACAACACACACCTTCCTAATAACGA

CGTATGTACTTCAGTAAAGATACACGATGAGGTACCGTATTCCAAACCTCGTTCCCCCGACTCGGACTGATGTAGTGAGCATTTTGA

AAGGTTTCTGTATGCATGTAACAAATGGAAACTCTTCTGAGAGAGAAGGTCTGAGAACCAGATGAATTATCGCCGTGTTGTTCCGCG

TGACGACGCAACAAAAAGCCTCCATGCCAACGCTCTCCTGCAACACAATCAGCTGAACTCCCAGCATGCTTTGCTTCTGCTGTTGCT

CAGCCAGCCGTGGAAATGTCTTGTTACCAAAACACACAAGCGTGAGCGGCGTCTCTGACCTCCACTGTGGCATTAGTTCTGCTACAA

TGGAAAGACTCTGTGGGAAGCCCCCCCTCCCCCCTCCCCTCAGTCAGTGGTGCTACTGATCACTGCTGCCAGAATCCTCACAACCGA

ACCTTAGTTGGGGTTATGAGGTCACAAGTCAAAGGACCCCCCTCGCTACCCAAACCGAGGCTGAATACCCCGTCCTGGTTTAACACG

AGTCTCAC***CGTCCTGGATCCTCCCGGCTAGGGCCTCGGAGCTGAAGCCAGCGAACACCACGGTGT***

**Exon 3**

***GGACCGCCCCGATACGGGCGCACGCTAACATGGACGCCACCGCCAGCGGGGACACCGGCATGT***

***AGACGGCAACCCTGTCCCCCTTTTGGATCCCGTGGCTCTTGAGGGTGTTGGCGAGGCGGCAGG***

***TTGTGTCCAGCAGCTCTCT***GGGGAGCAGACGGTCGGTCACTCGCTGATGGAGATGGTCGGAGTATTAATAAGCAAACATC

ATGTCAGCAGAGCAATGCAGTGATTTCACAGCCAATGAGACTCAGAGCAGCTACTTGTTAGTCGATTCTTTTTATGAGAAAACAGCA

ACTTCTGCAGATTGAGTCACTCACTTATGACATAAAGCCCGTTTCCTCCTCCTCTCGGTTATTGTGGTTAATCCTTACCCGACGTGT

GTGACCCGTAAAACATCACATTCCTTCATGTTTATGGTCGGAGAACGTAAAGCACATAAAGATGAGACCAGAGGTTTCTTTCCTCAA

GCTGTCCTTGTTTGTTAGGCTGGAGGGGGGGGGTCCATAC***CTGTAGGTCACCTTCACCTCTGTACCAGGCTCGT***

**Exon 2**

***CCCGCTCCCAAATCAGAGCCACTCGGTCGGGATGTGTCTCCACGTGGACGTCCAGGCAGTTCA***

CTGCAGCACAGAGGGAACCAGGCTATGTAATGTGATATGTGTAATGTGATATGTGTAATGTAATATATGTAATGTAATGCGATATAT

GTAATGTGATATATGTAATGTAATATATGTAATGTGATATATGTAATGTGATATATGTAATGTGATATATGTAATGTAATATATGTA

ATGTGATATATGTAATGTAATATATGTAATGTGATATATGTAATGCGATATATGTAATGTGATATATGTAATGTAATATATGTAATG

TAATATATGTAATGTAATATATGTAATGTAATATATGTAATGTAATATATCTAATGTAATGCGATATATGTAATGTGATATATGCAA

TGTAATGTGATATATGTAATGTAATGTGATATATGTAATGTAATATATTTAATGTAGAATTTAAGGGAACTGCCTTCATTCACTCGA

TGATTAAAGATTAATCCTGAAATGAGCTTAAAGCTGAAGACTTGTTTTTTGTTTTTGTCCTTTATGTAACTCAGAAATCCCCCCCCC

CCCCCCTCCACACACCCACAGACAAACACCGCATCTATTGGAAGGTAAAGGGATTACTACCAGGATCAGAACAACAACTGGGGGGGG

GGGTCTATTTACTGGTAATCAACAGATTAGTATTTCAATTTTACTTCATTTCTTCTACACTGTGGTATTATTTATTTGTAGTGTCCG

TGTGCTACTTCAGTGTATTAGTACTTTGTAATAGAGTATCTCTTCTGTGTATTAGTACTTTGTAATAGAGTATCTCTTGGGTGTACG

AGTACTTTGTAATAGAGTATTTCTTCTGTGTATTAGTACTTTGTAATAGAGTATCTCTTCAGTGTATGAGTACTTTGTAATAGAGTA

TCTCTTCAGTGTATTATTAGTACTTTGTAATAGAGTATGTCTTCAGTGTATTAGTACTTTGTAATAGAGTATCTCTTCAGTGTATGA

GTACTTTGTAATAGAGTATCTCTTCAGTGTATTAGTACTTTGTAATAGAGTATCTCCTCAGTGTATTAGTACTTTGTAATAGAGTAT

CTCTTCAGTGTATTAGTACTTTGTAATAGAGTATCTCCTCAGTGTATTAGTACTTTGTAATAGAGTATCTCTTCAGTGTATTAGTAC

TTTGTAATAGAGTATCTCCTCAGTGTATTAGTACTTTGTAATAGAGTATCTCCTCAGTGTATTAGTACTTTGTAATAGAGTATCTCT

TCAGTGTATTAGTACTTTGTAATAGAGTATCTCCTCAGTGTATTAGTACTTTGTAATAGAGTATCTCTTCAGTGTATTAGTACTTTG

TAATAGAGTATCTCTTCAGTGTATTAGTACTTTGTAATAGAGTATCTCTTCAGTGTATTAGTACTTTGTAATGGAGTATCTCCTCAG

TGTATTAGTACTTTGTAATAGAGTATCTCTTCAGTGTATTAGTACTTTGTAATAGAGTATCTCTTCAGTGTATTAGTACTTTGTAAT

AGAGTATCTCTTCAGTGTATTAGTACTTTGTAATAGAGTATCTCTCGGTGCGACGGAGGGTGAGTGTTGTCCTCCAGTGATCCCAGT

TCCTGCTAACCTGCGTGTCCTGGCAGGACTGTGTCTCCAGTGTGTGAGTGTTGGACACTGATGGACAGGTGTCTCTGTCTGGACAGG

TGTCTCTGTCTGGACAGGTGTCTCTGTCTGGTCTCTGCGCTCATCGACCGGTGAATGATGACACTTTTTACCATTTACCAGTCAGTG

TGGGATTAGATTCACTTCTAACAGTATTTCTTCCCTGTGGTATTTGTACTAGCATTCTTACAGTAGTACTTTCTTCTGGTAGTTTTC

**Exon 1**

TTCCTCAGATAACGGACACGGTGAAACTCGGATACCGACATCGCGCCGTTTCTCAC***CGGACACGTTGATCTTTCCGCC***
***CAGGAACCAGGCGATCCTCCCGGTGCTGAGGTCGCAGTCCCGGACCCGGTGGAACGGTTCGGA***
***CCAGCGCAGTCTGTCGGCGGCGGCAGAACCCCAAAACTGATCCGGATCCACGATCGACAGCCG***
***GTACAGATCCCGGTGGGACAT***CCGCGACAGCCGGGAGCTCCAGGTGCCCGGCGCGGGGAGAGCTGGCACGGGGGACCTG
TGCTGTCCCGGTCCGGCCCGGGTCGGAACCAGTACCGGGGTTCCGCAGCGACCCAGCAGGGTCGCGGAGAGTCTGGCGCTCCGGCGG
GTCTGAGCCGCCATGTTGATCAGCTGCTCAGAAGACCCCGTCCCTCTCAGGTGAGCTCAGGTGAGGCTGCGGCGGTACACGGTCACG
TGAGCGTCTGAGCCAATGAGCCGCTGATGAAGCTGAAGGACAAACTGCTGACGCGGTTTGAAATAAATCATTAATTAATAATAATAA
CAGAAATAATACTAGTGTAACCCTAGTATTATTAGTCCAATCCGTAATCATTTATATAACTACTAATACTACTTTAATACCAGTACT
GCCCATACAAATTCAAACAATCCTACTTATAGTCCTACTACTACTACTACTAGTTATACTGCTACTACTACCACCCTTACTACTGAT
ACTACTCATACATCCACACAGAGTAAGGACTCTTACTCCTCCTACTTATGCTAATACTTATACTACTAATAATACTACCAGAGTAAA
GGTTAGGCTAAAACCAGCAACTGGATAGCTTAGCATAAAGACTTGGAACTGCTAGCGGTTTGTAGGTAATTTTTTATAAAGTAAAGT
TCATAAAGTTCATCATTAATCCACACATGAAAGATGGACAATGTTACTAAATGTAGACAAAACATCGAGCAGCTTAAGTTAGAATAC
TTCTATAATACTATAGTATACTATAGTATAATACTATATAATATATAATAATCATCTGTATCTCCATCCATCCATCTTCAACTCCTT
ATCCGGGTCGGGTCGCGGGGCAACAGCTCCAGCAGGGGACCCAAACTTCCCTGTCCCACATCGACCAGCTCTGACTCCCCAGGCGTC
CCCAGGCCAGTGCAGAGACATAATCTGTCCACCTGGTCCTGGGTCTTCACATCCTCACCAGATGATCAAACCACCTCAACCCGAAGG
ACTCCGAGCTCCTCACGGAGCTTCTCACTCCATCTCTAAGAGGAAACCATTTGGGCCGCCTGTACCCGTGACCCCGTTCTTTCAGTC
ATTCGGTTCAGGATCTCACCGGGTTAAATATGATACTGTTCTATAAAAACAACTGTGTGATGAGAGATATTGACTCTAAACGCACAA
TAAACGCACACAGGAAACACGCAGCAGTGGAGCTCAGACTCTGGAGTCTCTTGGAAGTAAATATGAATATGAGGCTTGAAGTCATAA
CAACATGTAACCGCTGTGGTCAGATGCTTGTTGTGCTGCTGCTGCTCAGTGACCTCTGACCTCAGATCTGCCATGTCTCCATCAGAT
TCTTTAATCTGACGGAGCTGCAGCGTAATCGCTTGTAATTCCCACTTTACTGCCATCACATGATAATTGATGCTAGTCTGTGTAATC
CAGCTGACCGGAGGTCAGATTAAGGACGTATACGCCTTTGTATTTCCACCATAGATTTAATCATCTTTGTGACTTTAATGATGCTGA
TGGCTCATTGTTGATGTTGTTCTCCTCATGCTGTTTACCTCTACGACGTGGTCAGCGGTCGCTCTTGCTGCAGACGGAGCGGAGCGG
CCTGTTGGTTGGCGGCCTGCGTGAAGCCCTTCCCGCCTCCCGGTGGAGCAGCTGCTGCCGCGTCAGCTGGCCGGCGTGGAAGAAGCT
GTTCCCGCAGACGAAGGTCTTCTGGAAGTGGGTGGTCCGGTGGGTCTTTAGCCCGCTGGCGCCACAGGAGCCCTTCCTGCGGCGTCG
TAGATCCTCCAGGTCTCCTGTATGCTCGACCGTCTTCAGGACTTCTCCATCACACAGCCGGTCCACTAGAGAGCTCAGCATCTTAAA
TCAGAGCATTTTTACATGGAGTTTGGTGTAACAACGTGTCAGTTAGCAAATAGCCAGTATAATAAAATAACAGAGCAGAACTTATTA
AATTAACTTTAATGATATTGAGTGAAATCAGAGTGAAAGTATGTTCAGTGAGGTGAAATAGGTATTAATTAAAATTAAAATATTAGT
GCAATGGTTATGGATAATGGTCCAGTTTTAAATGGCGGTCATGTGCACATTTCAAAATGAAATGCACCTCCATTCATGGCGTTACGG
TTTAAAGGCGCTGTTAAGCGTTTCCTCTCTGCGGCGTCTTTAAAGCAGATACTTTGAATGCTACATTCTGATTCCTCCTCCCCGCCC
AGCCACAGCCTGGCACTTGGTCAGAGAAGTGGAGGCAACAAACCCCAGGTGAGGGGGAGAAGACCAGGAGGTCCTTGTGAACATCAA
CCATGTGAAAGGAGAATGTAGGCACCATAACGCCATGAATAGAGGTGCACTTCAGTTTGACACCTGCTCCGGTTTAGGGGGTAATTG
TTGCAGTGTCGGTCTATGCAGGAATGTCTGTGTCTGTCCCTGCAGCAGGTGGTCCAATATGGACAGTAGTTCCTTCAGTGTCCTCCT
CCACCTGGATACTCCTGTTAGCAGCTACCAGCTCCAGCTCACCTGTTTAGTCACACAACAGACTTGCGGTGACGCGCTGAGTATTTC
CCGCCACACCGATTCGATCGCGAGGAGAAGTTCATTTTTAAAACAGGTAACGTGGGACACGATGAAGAGGTTCTGCACGTAGCAGCC
GAGTAGCTTAGCAGAGAAAGACGAAGGAAGTGACGAACATTTCTGCCTCCAAATTAAAACCACCTTATTTTTGATGATGTTCTTATG
AGCGTTGACCCGTAATAAGCTCAACGGAGCAGCATTTACATTGTTTTATAAAAAGGATTCTCACATGATTGAAATATTTAATGTCTT
TATAGTTATTTGAAGCTGCTGTAAACCTGCATTATAGTAAACAAATACATAGAAGCTTCAATAATCGAGGTTTAAACATTTTATGAC
AATTGTCTAATGAAACGTCCATTAAACATTTCCTTTCTCGACTGAAGTATCAACACGTATAAATGTATAATAATAAGTGATTAATGA
CGTTTATAATTAAACGTAAATTCTGCAGTGTTTGTGTGCCGTTGAAAAATCCCAGTGATTCCAAAAATACTGTTTTTTATTCACCAA
TCGGCTGCTGTACATGTTGATGATAATAATCCCAATAATGATAGTAATAATAATAACCACCATTTAAAATGTCACATTTGAATGACA

```
TCACTCACAGAGCTTCAGTCAGTGACAAGCAGTCAAACTATAAAACACGATGGGAAGTAAACAGGAAATAGCCTCTTAACCCTTTGC
CCCAAAACACGGATATAAAACATACTCTCAGTACAATACAATAATAAACACACAGGTACGTCTTCCTGAAACAGGTTAAAGGTCAAC
TCTCAGTCCGATTCCTTGGGGGCGACCTGGCGGCGGCAGAGTTAGCATGCTAATGTGTCTGCGTGCTAACCAAGAGCAGCATTTCAA
TTGTAGCCATTTTATTTTGCCTCGTTCTCTCTTTGGCCCGTCGGTCTGCGAACAGCCAATCAGATCCCTCCGTTCGCGCCTCCGGCC
TTTAGCTGCGTTCGTCCGATTCGCTGCCGCTGTGCCGCGAAGAGGCGGGGCTTACGGAGCTCCTGGGCTCCTGCTGCTGCTTAGAGG
AGCTGGATAGGTCGATGGCCACCTCCTCCTCCACCTCCTCCTCCTCCACCCCCAGCGGCTCCTCCTGCCGAGCCTCGGCGGGGCTGC
GTCTCTCTGGGGCGTCGCAGAACGAGTCGGGTCGGGACGGGTCCGCCTCCCGAGGACTGGGACACTTTTTGGTTATTTCTTGAGACT
TCTTGTGCATTCCTGTGGGGACGGAGACGGAGAGTCAGAGATACCACATGTGTGTTCTTAGTTCAGTGTAATATACAATGTTATATA
TATATATACACGCATACAAATGTGTACTTTAGTAGCATCCACCTGACCAGCTGGCGTTAAGACTCACTGGATTTATCTGAATTTTTT
TTTTGAGTCATGTGATCAAATTCAGAGAACTATTTTAATAAACATACATATTTTCAATTTAACTTAGTTATCTGCAGATGGAAACAC
AAGTGTATTCAAACAGTGTGTGCGTGTGTGTGCGTGTGCGTGCGTGCGTGCGTGCTACTGACGTGCATGCGGCTCACCGAGCAGCCA
TGGAGCGCAGGACCCCATCATGCCGTTCTTGGCCGAGTTGAGGATGGACTCGGGCAGCGGGATGGAGTGTCTCACCATCGCCCCGTA
CAGGCCGTACTCCGCCATCACGCTGCTGCGGCCCCAGCACTTCTCCCGCTTCCTCCACTTGGCCCGCCGGTTCTGGAACCAGACCTG
CACCACGCGCACCGGGTCAGACGTCCACTACGCCCACTTAATCATCAACGTGATTATTTTCTATCATTATTATACTTCACCTTCATA
TTACTGTCGATCCTCTTTTTTAACATTCTTATAGTTTTCACTTGTATTATCTTTCTAATTTCTCTCGTTATTATTATTATCACAATT
ATAGTTGGAATTATTATATTTCATTGATTGCTCGATTTTTTATTTTCTTCTTATTAAATATTATTTTTAAATGATCAACAGACTGTG
TTCCATTAATATAATAATAATAATAATAATATAAGTTGACAGATGAATGGATGAATGTGTCTACCTGTATCCGGTCTTCAGGGAGCT
CTGTCTTCATGGCCAGCATCTCTCGGGCGTACACGTCCGGGTAGTGGGCCTCGTGGAAGGCCTTCTCCAGCTCCTCCAGCTGGTGGG
ACGTGAACACCGTCCTGAGGACAGACAGACGACCCGCATGTTTAAACCTTCCACCTCTTAAATTCTTCTTCCCCGTGAAGGGCTCGT
AGTTCCTGTGTAGATGTGAGCGTTTCAGGACGGAGGGTCATGTGTCCACACTGTAAAGCCCTCTGAGGCCAATCATTTAATTCATGT
AATTTGCCATTTTGGGCTCTACAAAATAAAATGAATTGAATTGAATAAATAAATGAAAACATTTTAAAAAACAAAAACAATCGAATA
TGAAATATTAATATAAAGTACTGTATCGTGCCTTAAGTTCATAAATGTATATGGCTCTGTATAATAAATATCCTACAGATATGAGGC
TCACCTGTGTCGCCTCTTCTTCCTCTTTTGAGAGTTCGCCGAGTTTTTGGAGTCGTTGCGGTCAGAGAGACACTCTTCATCTGCAGA
AAACAGGAAACACGTCATCAGCCGTCACTCAACGTCAGCGTGAAAACGCTCACTGACTGATTCATCGATCTCGATTCAATCGAAAAG
AAATAGGCACATTTATTAATTATGAGGAATTTAACCCCCACAGGAATAAAAGTTTAAACAATGCTGAAACATCGTATTATTTTAGGA
TCCAACCGAACCGAATCCAGCCAGAAACAACATCCAGCGTTCCGCATCACAATACAACATATTCATCTGAAAAAGAAATAACGAGGA
CAACAAATGAATGCAGCTGCTGAAAGCTCGAACCGAGGCGGACTATTATTATTATTATTATTAATAGAAAAATATCCTTATAACAAT
AAAGCGATTTTAAATCGCCTCAGTGGATTTGTTAAAACGAACTTGCTGGTTTTAGAGTTTAATTGAACAAAAGTGAGGAAGCTTTAA
AAGTGGAACGGAACCGAGGGGCGGGGGCCGCTGTACCGGAGTAGGCGTCTCTCTGGGCCTCCAGGTTCTGCAGGTAGTGGCTGTCGG
CCCGGGCCTGTAACAGCGGCAGGTGTCCCGGCAGGAAGCAGGGCGCTCCGGCGGGCTGCTGGGCCGCCAGGTTGCACAGGAAGCCGA
GGCCGAGCGGCAGTGACCCCCCCCGGGAAGGAAACCCCGCCGAGTCCGCCTCCGACCCCGTGCGGGTCCGAACCCGGGCCCCGAGCC
CGCCACAGAAAGTCCCGGAACCTGGGGGCCCGCCGGGCTGCTGCCGCGCCCTGCAGGTCCGACTCCAGGCCAAGCAGGTCGGTGATG
GCGAAGCCTTTTGACCCGGAACCCGGTGCCGTTCAGTCGGGATTTATCGATCCCGAATCCCGCAGATAACATCTTGACCTTGGGCTT
CTCTTCCAGCTCCTCCCGGGCCGTCATGCTGCTCATCAGGGGGCTGAAGTCCGGTCCTTGTGGTCTGAGATAAAGAGGAGGAGGAGG
ATGGACAGACCGGGACAGTCCGGCTGGCGGGCTCCATTTATCGGGCCCCGGTCTCCTGGATCACTGCCACATCCAATCAGAACCGGG
AGTGGCAGAGGGGGCGTGTTCCTCCGGTCTCCTCTCAATCCCAGAGGATTAATTGCCACCAATTTCCCTCCAATCCGATCAGGATTT
CTCCATGAGGACAGAGAAGGGGACTTTTAATTTACGTCGTTTATCTTTGATCAAAGAGACCAAGAGAGTTCAAACTAATCAATTACA
TGATGAGAAACAAATCCATTATATTATGATTTAATACACAAATCAATGAAAATAATGCTCAAATCGATTATCTAACGAATCACAAAA
CAATGATGATAGAAAGTAAAAAACAAATTTATTGAACCATCAATCAAAAAACAAAAAAATGATATGATGAAGAAATATACAAATCCA
GTGAATTATGAATTAATACACAAATCTATTATATTATTACACAAATCTATTCCACTAGTAATTTGTATTATTGTGAATATAAAACAC
CAATCTAATGTAAATATGGCTATGCATTAATGCATTAATCACTCACATTGTGCATTTGGAATAAATTCATTTCTGATTTTTCCCACA
```

```
TGTGCGGGTTAATATCAATAATTTATTAAAAGAGTAAATAACAAGAACATGTCAGTTAATTATATTTCACAATATCGCATGGTGGTG
CATGTTTATTTGGTAATACAGTATCATAAGATTAATTTTTCGAAATAAAAACTAATTAAATAAAAACATTTAAGTCTGATCAAATAT
TTTAAAAGTGAACTTTCAACCTTTGTCCAAGTAAACTGTCTCTATGTGTAACATAATAATTAAACACATAAATAAATGCACACAAGA
AGAAACCAGATGAAATATGAACACACACACACACACACACACACACACTTGTTCATTTGCAGACGTAAAAAAACAATGAAGCAGCTT
CAGAGTTTATTCAGATTATAAATGTTTTAAATCAGAACATCAGTTGGACTCATGTAATCTGAAAAGAAGGTGCTGAGGTTGGATTAA
ACTGTGTGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAGCGCCCTCCACAGAATCTTATGTTCATTTC
TGAAACAATGTGGATTACCTTGATATGAGGAAGGAGATTGATTGATTATGTGCTTCTGAATAATCTCAGATTAACTTTTTTTATTCT
TGTTCCTCTGATCTTTTTATCCTCGTCCTCACGTCGCTCTCCGTCCTTCGTCTCTTTATCGTTTTATTTCTGCCTTTTGTAATTATT
CTGCAGCGACTGTCATCACTGATCTGAGATCAGGTTTAGAACATCATCAGTATAAATATAGTGAAGTGTTGATCAGTCAAATTACAT
AATTTCCTTCTATGTTTCTGGATTTATTCCTTGAGAGGCAAATACAAAAATACCCAATACTCTATTCCCATACAAGCCACAGTTAAT
AATTCATATCTGTGTATAATTATAATAATAATACGAATATATATATATATGAATTTGATGAGTTGTGGTTCATAAACTGTGTCGGTT
TAGTATTAATAATAATAATAATAGTACATATGATATTGATGAGGTGTGGTTAATAATAAACATATGTGTTTATTAATAACCATGACC
AGTTTATTCATCACAATAACCAGTGTTTATTGACACAAATATACTGTTTGTAAATTCCACTAACCAGTTTATTAATAACAATAAGAA
GTTTATTAATAACAATGACCAGTGTTTAATAACAATACACAGTTTATTAATTCCAATAATCAGTGTTTATTAATAACAATGACCAGT
GTTTTTTAATTCCAATAACCAGTTTATCAATAACAATAAGCAGTTTATTAATAACAATAACCAGTGATTCGTAGTGGCAATAACCAG
TGTTTATTAGTAACAATAACTAGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAACAATAAGCGGTGTTTATTAGTAACAATA
ACCAGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAGCAATAACCAGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAA
CATTAAGCAGTGTTTATTAGTAGCAATAACCAGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAACATTAAGCAGTGTTTATT
AGTAACAATAAGCAGTGTTTATTAGTAGCAATAAGCAGTGTTTATTAGTAGCAATAAGCAGTGTTTATTAGTAACAATAACCAGTGT
TTATTAGTAACAATAACCAGTGTTTATTAGTAGCAATAAGCAGTGTTTATTAGTAGCAATAAGCAGTGTTTATTAGTAGCAATAAGC
AGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAACAATAAGCGGTGTTTATTAGTAACAATAACCAGTGTTTATTAGTAACAA
TAACCAGTGTTTATTAGTAACAATAACCAGTGTTTATTAGTAGCAATAAGCAGTGTTTATTAGTAACAATAACCAGTGTTTATTAGT
AACAATAACCAGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAACAATAACCAGTGTTTATTAGTAACAATAACCAGTGTTTA
TTAGTAACAATAAGCAGTGTTTATTAGTAACAATAACCAGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAACAATAACCAGT
GTTTATTAGTAGCAATAAGCAGTGTTTATTAGTAACAATAACCAGTGTTTATTAGTAACAATAAGCAGTGTTTATTAGTAACAATAA
GCAGTGTTTATTAGTAGCAATAAGCAGTATTTATTAGTAACAATAACCAGTGTTTATTAGTAACAATAACCAGTGTTTATTAGTAGC
AATAAGCAGTGTTTATTAGTAGCATAAGCAGTGTTATTAGTAGCAATAAGCAGTGTTTATTAGTAGCAATAACCAGTGTTTATTAGT
AACAATAAGCAGTGTTTATTAGTAGCAATAAGCAGCTGGCGGGTGCTGATGACACTAATTAGACCTGCTGCCTCAGTAAGCGATAGC
AGTAATTCAGATGTTGCTGCTCCCGCCAGTTGTTGTTCATTGTTTTCCCTCACTATTGTTTACTGGCAACGGCAACAATAGTCTGGC
CACGCGTGCCCCCCCCCCGACCCCCCACCAGACTCCACCTACTGCCCTCTGTGGCCCACTAATCCCTTACATAGAATGCACGTTTAT
GCAAATAACCCCGGTTAGATTA
```

## >GacACSS1A_predicted protein sequence

```
MSHRDLYRLSIVDPDQFWGSAAADRLRWSEPFHRVRDCDLSTGRIAWFLGGKINVSVNCLDVH
VETHPDRVALIWERDEPGTEVKVTYRELLDTTCRLANTLKSHGIQKGDRVAVYMPVSPLAVAS
MLACARIGAVHTVVFAGFSSEALAGRIQDAQCKAVVTFNQGVRGGRLIDLKATVDSAVKNCPT
VQHVFVAQRTENPAVMGQLDVPLEEAMSSQSPVCPAEPLDSEDLLFLLYTSGSTGKPKGIVHT
QAGYLLYTSLTHQYVFDHQDGDVFGCVADVGWITGHSYVVYGPLCNGSTTVLFESTPVYPDPG
RYWETVQRLKISQFYGAPTALRLLLKYDESFVKKYDRSSLRTLGSVGEPINHEAWHWFHSVVG
EGRCPVVDTWWQTETGGVCIAPRPAEEGAAIVPAMAMRPFFGIQPSLLGEKGEVLLGDGVGGA
```

LCISQSWPGMARGIHGDQQRFLEAYFKPFPGYYFTGDGALRSADGFYQITGRMDDVINVSGHR
LGTAEIEDALDEHPAVPEAAVIGFPHEIKGEVPFAFVVLKEDLGVDPLLVLQQLRDLVSTKIA
RYAVPEHFLVVKRLPKTRSGKIMRRILRKVAMETTGDLGDVSTLDDPSVVMEIIEAHKQYRTQ
RGGDK

## Supplementary Table 1

| Gene | Motif I | | Motif II | | | | | | | | Motif III | Motif V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (10aa) | | (56 a.a) | | | | | | | | (5 aa) | (10aa) |
| HsaACSS1A | YTSGSTG | M | TGDGAYRT | E | G | G | Y | Y | Q | WWQTE | PKTRSGKVMR |
| | P | K | I | T | G | R | M | D | D | | |
| | | | V | I | N | I | S | G | H | | |
| | | | R | L | G | T | A | E | I | | |
| | | | E | | | | | | | | |
| MmuACSS1A | ....... | T | .....H.. | . | . | . | . | . | . | ..... | .......... |
| | . | . | . | . | . | . | . | . | . | | |
| | | | . | . | . | . | . | . | . | | |
| | | | . | . | . | . | . | . | . | | |
| | | | . | | | | | | | | |
| XtrACSS1A | ...... | . | .......S | N | D | . | . | . | . | ..... | ...... . |
| | K | . | . | . | . | . | . | . | . | | I . |
| | . | | . | . | . | . | . | . | . | | . |
| | | | . | . | . | . | . | . | . | | |
| | | | . | | | | | | | | |
| GgaACSS1A | ...... | . | ........ | K | E | . | . | . | . | ..... | ...... . |
| | K | . | . | . | . | . | . | . | . | | I . |
| | . | | . | . | . | . | . | . | . | | . |
| | | | . | . | . | . | . | . | . | | |
| | | | . | | | | | | | | |

## Supplementary Figure 2

```
                                   ↓    ↓      ↓                        ↓      ↓↓    ↓
HsaACSS1A  ELLETTCRLANTLKRHGVHRGDRVAIYMPVSPLAVAAMLACARIGAVHTVIFAGFSAESLAGRINDAKCKVVITFNQGLRGGRVVELKKIVDEAVKHCPTVQHVLVAHRTDNKVHMGDLDVPLEQ
MmuACSS1A  ..............................................I...V...................A.........................S..............T..P..S..I....
XtrACSS1A  ...N....I.....KY.IQ......................T........V................C..A............TT....T......N..SIKN.F..Q..E.E.P..KI.I...E
GgaACSS1A  ...DL.........KY.IQK..K.....S....S................V.............M.SE..A...Y...V....II...TT......N..S.K..F..Q.....TQ.........E
AcaACSS1A  ...DM.........KN.ITK.........M....................V...............SE..A...C........II....T.....QK..SIKC.F..Q..A..I..RNQ.IL..E
DreACSS1A  ....M.........S...Q..E..........M.................V.....S.A.....Q..Q..F...C...V.....FD..ST..K...S..S.R..F..K..N.S.P..K..I...E
TniACSS1A  ..................QK.....V...........L............V.....S.A.....Q.....A...C.EAV....LIP..AT..A..RS....R..F.SQ..EKQCV..E......E
TruACSS1A  ..............S..IKK..T..............L............V.....S.A.....Q.....A...C.E.V....LIP..AT..A.L.S....R..F.SQ..EKQCM..EM.....E
CmiACSS1A  ........I.........QK..T.........V.............I.S.......SQA.ST.....Q.TI...C..SV....II....T......M....KR.F.EQ..E...Q.N.I.I...E
GgaACSS1B  ....L....S.....Q..K.....T....PC.....S...........A.V..........D..R..QSET...V.........K......T..Q...Q..G.KR...SM...SQLS.TA...L..E
AcaACSS1B  ....K..........Q..KK....V....PC.M..MS.........I...V.....SVA..D..Q...SET...V..SI...K.....QT..Q.I.M..M.KR.F.SK...A..S.SAV.I....
GacACSS1B  Q...M....G.L.R.R..K...C.T....PC.....S.M.......A.N.V.......A..E..R..QSST...A...V...K.T....T..A..QS..A.RQ.F..M..E.P.A.TAR..AM.E
OlaACSS1B  ....M...VG.L......RK..C.T....TC.....S.........A.N.V......DA..E..R..QSSI...M...V...K.T....T..T..QS....R..F..M..ETL.P.T.R..LMDE
TruACSS1B  ...DM....G.L.R.R..K...C.T....PC.M...S.........A.N.V.......A.SE..R..QSST...V...V...KL.....T.....QS.....Q.F..T..E.P.L.AAR..A.DE
TniACSS1B  ....M....G.L.R.R..K...C.T....SC.M...S.........A.N.V.......A.SE..R..QSST...V...V...K......T.....RS.SS..Q.F..M..EKPAE.TAV..A.DE
CmiACCs1B  Q...V..........R..R.....TL.L.A..I...S.............V......DA..N..Q..QSET...V..A......I....T......S..SIKQ....T..E...P.S.I.....E
```

**Supplementary Fig. 2**. Partial protein alignment of ACSS1 sequences. Arrows indicate diagnostic amino acids between 1A and 1B paralogues. Hsa—*H. sapiens*, Mmu—*M. musculus*, Gga—*G. gallus*, Aca—*A. carolinensis*, Xtr—*X. tropicalis*, Tni—*T. nigroviridis*, Tru—*T. rubripe,* Ola- *O. Latipes*, Cmi- *C. milii*

## III.2 Diversity and history of the long-chain acyl-CoA synthetase (*Acsl*) gene family in vertebrates

Mónica Lopes-Marques, Isabel Cunha, Maria Armanda Reis-Henriques, Miguel M. Santos,

L. Filipe C. Castro

**BMC Evolutionary Biology**

(DOI: 10.1186/1471-2148-13-271)

Lopes-Marques *et al. BMC Evolutionary Biology* 2013, **13**:271
http://www.biomedcentral.com/1471-2148/13/271

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

# Diversity and history of the long-chain acyl-CoA synthetase (*Acsl*) gene family in vertebrates

Mónica Lopes-Marques[1,2], Isabel Cunha[1], Maria Armanda Reis-Henriques[1], Miguel M Santos[1,3]
and L Filipe C Castro[1*]

## Abstract

**Background:** Fatty acids, a considerable fraction of lipid molecules, participate in fundamental physiological processes. They undergo activation into their corresponding CoA esters for oxidation or esterification into complex lipids (e.g. triglycerides, phospholipids and cholesterol esters), a process that is carried out by acyl-CoA synthases (ACS). Here we analyze the evolution of the gene family encoding for the long-chain acyl-CoA synthetases (*Acsl*) in vertebrates.

**Results:** By means of phylogenetics and comparative genomics we show that genome duplications (2R) generated the diversity of *Acsl* genes in extant vertebrate lineages. In the vertebrate ancestor two separate genes originated the current *Acsl*1/5/6 and the *Acsl*3/4 gene families, and the extra gene duplicates in teleosts are a consequence of the teleost specific third round of genome duplication (3R). Moreover, the diversity of *Acsl* family members is broader than anticipated. Our strategy uncovered a novel uncharacterized *Acsl*-like gene found in teleosts, spotted gar, coelacanth and possibly lamprey, which we designate *Acsl2*. The detailed analysis of the *Acsl2* teleost gene *locus* strongly supports the conclusion that it corresponds to a retained 2R paralogue, lost in tetrapods.

**Conclusions:** We provide here the first evolutionary analysis of the *Acsl* gene family in vertebrates, showing the specific contribution of 2R/3R to the diversity of this gene family. We find also that the division of ACSL enzymes into two groups predates at least the emergence of deuterostomes. Our study indicates that genome duplications significantly contributed to the elaboration of fatty acid activation metabolism in vertebrates.

**Keywords:** acyl-CoA long chain synthetase, Gene loss, Genome duplication, Differential paralogue retention, *Acsl2*

## Background

Two rounds of genome duplication (1R and 2R) have now been clearly established to have occurred in early vertebrate evolution [1], with a further round taking place in teleost ancestry (3R) [2]. Extra independent genome duplications have been determined in salmonids [3], and in ray-finned fish paddlefish [4]. These events have modeled the genomes of extant vertebrate lineages in terms of gene numbers and the overall genome architecture, contributing to the appearance of numerous innovations [5]. In addition to the increase in gene counts resulting from 1R/2R/3R, the comprehension and detection of gene loss processes in combination with the differential retention of paralogues poses important challenges to enlighten vertebrate evolution [6-9].

Fatty acids (FA) are a particularly important category of lipid molecules, being a considerable source of energy and a significant component of bio-membranes. FA metabolism involves among others, processes such as hydrolysis, beta-oxidation, synthesis, esterification and activation. The later comprises a two-step, ATP dependent reaction, with the formation of an intermediate acyl-AMP which is then converted to acyl-CoA. Acyl-CoA synthetases (ACSs) are the key enzymes responsible for this fundamental initial step in lipid metabolism. They can act on non-polar hydrophobic FA substrates and convert them into water-soluble products (acyl-CoAs), which are then integrated into metabolic pathways such as oxidation of acyl-CoAs to obtain ATP, storage in the form of triglycerides (TGA) or use as building blocks for other lipid molecules. The human genome contains 26 ACS genes divided into 6 distinct families: Short-chain ACS family (ACSS); Medium-chain ACS family (ACSM); Long-chain ACS family (ACSL); Very long-chain ACS family (ACSVL), Bubblegum ACS family

* Correspondence: filipe.castro@ciimar.up.pt
[1]CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, CIMAR Associate Laboratory, UPorto, University of Porto, Porto, Portugal
Full list of author information is available at the end of the article

(ACSBG) and ACS-Family (ACSF) [10-12]. This division reflects the chain length of their preferred substrate. ACSL enzymes play a paramount role in humans, since FAs with 12 to 20 carbons (C12-C20) are highly prevalent in the diet and are preferentially converted to acyl-CoA by these enzymes [12,13]. Further, several pathological conditions have been linked to ACSL enzymes such as inadequate lipid metabolism leading to diabetes [14], X-linked 63 mental retardation (MRX63-OMIM300387) [15,16] and cancer [17]. In mammals, previous studies identified five distinct *Acsl* genes, which were further organized into two separate groups (*Acsl1*, *Acsl5* and *Acsl6*) and (*Acsl3* and *Acsl4*) [10,12,18]. It is worth mentioning that the current *Acsl* gene nomenclature omits the "*Acsl2*" which was dropped since shortly after its identification it was found to be identical to *Acsl1* in human and additionally rodent "*Acsl2*" was also renamed *Acsl6* since it shared more identity with human *Acsl6* [18]. The advent of whole genome sequencing projects allowed the identification of *Acsl* genes in non-mammalian species, but no approach has been made in order to unravel the evolutionary history of this family [12,19]. Additionally, recent findings have illustrated the need to consider genome duplication processes (and gene loss) in combination with extensive species analysis for proper evolutionary insights regarding lipid metabolic gene networks to be drawn [20,21]. Moreover, non-mammalian species, such as the zebrafish, have been recently proposed as alternative models to study lipid metabolism [22]. Therefore, a comparative and phylogenetic approach involving a broader number of vertebrate species should shed light into the evolutionary history of *ACSL* enzymes and their function. In this study we demonstrate that genome duplications in stem vertebrate ancestry and the teleost specific genome duplication were instrumental in the generation of *Acsl* gene diversity. Moreover, we show that the variety of *Acsl* family members is broader than anticipated. Our strategy uncovered a novel uncharacterized *Acsl*-like gene found in teleosts and coelacanth, which we designate *Acsl2*. The detailed analysis of the *Acsl2* teleost gene *locus* strongly supports the suggestion that it corresponds to a retained paralogue, lost in other vertebrates classes ("*an ohnolog gone missing*"). Finally, we provide the first comparative transcription analysis between the human and zebrafish *Acsl* gene repertoire.

## Results

### ACSL gene repertoire in vertebrates

Human ACSL1, ACSL3, ACSL4, ACSL5 and ACSL6 sequences were used to perform Blastp searches and collect *Acsl*-like sequences from various available genomes. We analyzed a total of 21 species in order to include all major vertebrate lineages. Our database search determined the presence of five *ACSL* genes in humans, mouse, opossum, chicken, anole lizard, western clawed frog, and the coelacanth. In the spotted gar, an out-group of the teleost specific genome duplication [23], we found 5 sequences though 2 were partial (Additional file 1). Blast searches in teleost fish genomes hinted at a larger *Acsl* gene set, with nine hits in zebrafish, pufferfish, green spotted puffer and medaka and seven in stickleback. However, detailed sequence analysis suggested a number of inconsistent annotations in the Ensembl database. For example, we found three *Acsl1* gene annotations in medaka, (1-ENSORLG00000019563, 2-ENSORLG00000018806 and 3-ENSORLG00000008655), however, when aligning the DNA and amino acid sequence of the first two sequences we observe that they are identical (not shown). Given that the annotated *Acsl1* copy ENSORLG00000018806 is located within a contig that presents extensive regions that are poorly resolved we consider that this species presents 2 gene copies of *Acsl1* and select ENSORLG00000019563 and ENSORLG00000008655 for further studies. The green spotted pufferfish again shows three annotated copies of *Acsl1* with two of these (ENSTNIG00000000345 and ENSTNIG00000010115) located in the same scaffold with the same orientation and contiguously (Additional file 2). These annotations are partial sequences, one corresponding to the N-terminal and the other corresponding to the C-terminal of the protein. Here we assume that these annotations correspond to a single gene poorly assembled. Therefore we consider that the green spotted puffer presents two *Acsl1* genes and we use only the correctly annotated gene (ENSTNIG00000018054) for further analysis. Finally we find two annotated *Acsl1* genes in pufferfish (1-ENSTRUG00000017576 and 2-ENSTRUG00000001450) were the second gene corresponds to a partial sequence which was not used for further analysis. We investigated also the genomes of three Chondrichthyans, the elephant shark, catshark, and little skate. Our investigation identified 4 full sequences and several partial (Additional file 1).

Finally, the search in the lamprey genome resulted in four *Acsl*-like gene hits (1-ENSPMAG00000008135, 2-ENSPMAG00000004625, 3-ENSPMAG00000005099 and 4-ENSPMAG00000005133). Three of these correspond to partial sequences (449 residues) and were not used for further analysis. Finally, in the investigated invertebrate species, acorn worm and amphioxus, we recovered 3 *Acsl* sequences from acorn worm and 4 *Acsl* sequences from amphioxus. After clarifying all inconsistent gene annotations a set of ACSL sequences from various species were collected to perform phylogenetics (Additional file 3).

### Phylogenetics indicates vertebrate specific Acsl gene expansions

Preliminary phylogenetic analysis confirmed that *Acsl3* and *Acsl4* form a distinct group from *Acsl1*, *Acsl5* and

*Acsl6* as previously reported (data not shown) [10,12,18]. Thus, we have reconstructed each phylogeny separately (Figure 1 and Figure 2). In the Maximum likelihood analysis Figure 1A it is possible to observe that invertebrate sequences out-group four statistical well supported clades comprising *Acsl1*, *Acsl5*, *Acsl6* and an unidentified *Acsl* group. However, the exact phylogenetic relationships between each isoform are not statistically supported with the bootstrap analysis. In the Bayesian analysis (Figure 1B) we find again the invertebrate sequences out-grouping four statistically well supported vertebrate clades. The unidentified *Acsl* group is composed of teleost, rayfin fish and coelacanth sequences. In the Maximum likelihood analysis a lamprey sequence also groups with this novel clade (though weakly supported). We name this new gene lineage *Acsl2*. The overall tree branching pattern in Maximum likelihood and Bayesian analysis is indicative that the expansion of *Acsl1/5/6/novel* clade took place after the radiation of the vertebrate lineage approximately 500 million years ago, although independent gene expansions have taken place in amphioxus and the acorn worm (Figure 1A and B). We find representatives of *Acsl1/5/6* in all of the examined vertebrate species, with the exception of lamprey and chondrichthyans where the presence of partial sequences impedes a final conclusion regarding the full *Acsl* gene repertoire in these lineages (see Additional file 1). Nevertheless, this cannot be taken as an indication of gene loss due to the poor genome sequence coverage. The phylogenetic trees also indicate that *Acsl1* has specifically duplicated in the teleost lineage. Even though only medaka, zebrafish and stickleback present these duplicates, we antecipate that pufferfish and the green spotted pufferfish probably retain these two copies.

Regarding the *Acsl3* and *Acsl4* trees (Figure 2), both in the Maximum likelihood and Bayesian analysis we observe that the invertebrate *Acsl*-like sequences again out-group two well supported groups containing vertebrates sequences. Also, it is possible to recognize that all teleost species here analyzed present a lineage specific duplication of *Acsl3* (*Acsl3a* and *Acsl3b*). In zebrafish and cave fish we find an *Acsl4* duplicate; microsynteny analysis of this *locus* in zebrafish suggests that this extra gene copy is also a teleost specific 3R duplicate (Additional file 4). However, despite extensive database search, we did not

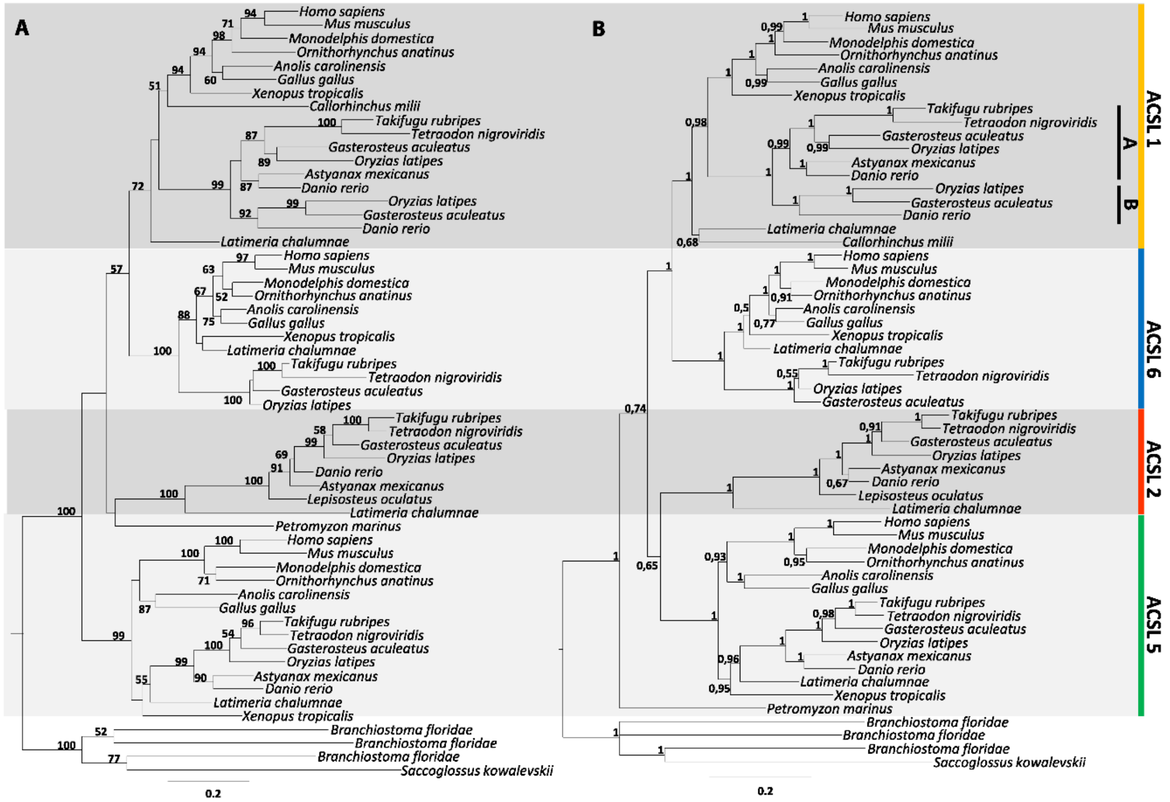


**Figure 1 Phylogenetic analysis of Acsl1, Acsl2, Acsl5, Acsl6 sequences. (A)** Maximum likelihood tree, numbers at nodes represent percentage of bootstrap values (only values above 50% are shown); **(B)** Bayesian tree, number at the branch nodes refer to posterior probabilities.

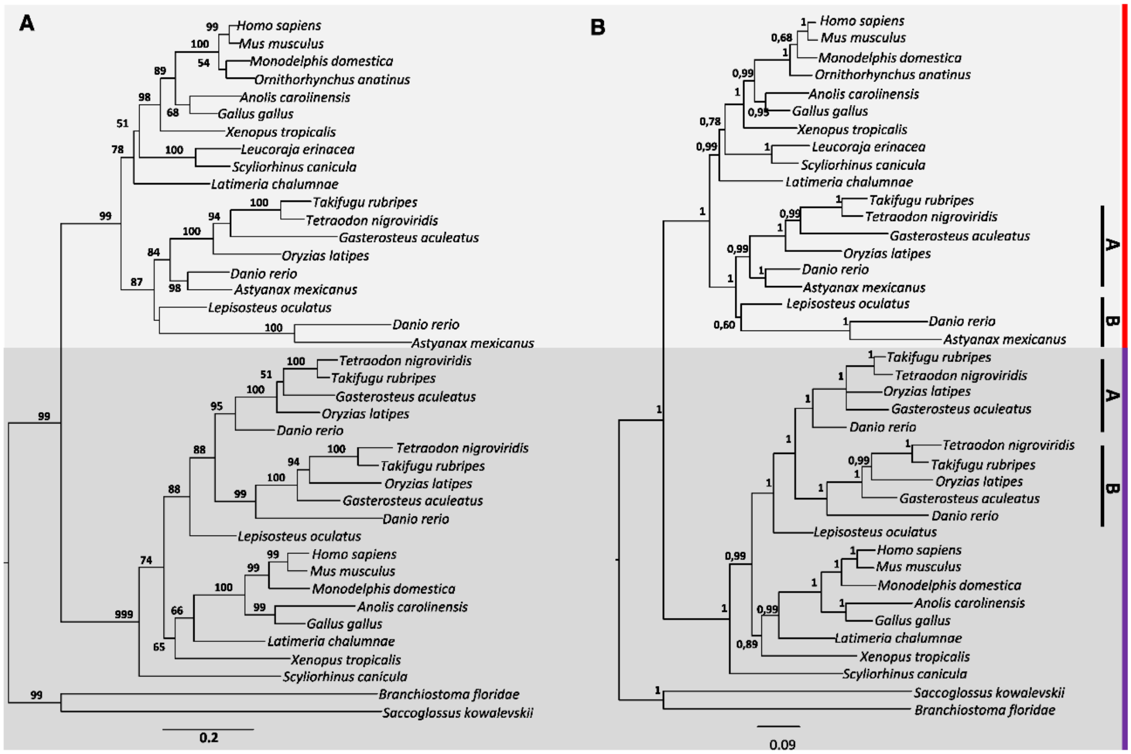


**Figure 2 Phylogenetic analysis of Acsl3 and Acsl4 sequences. (A)** Maximum likelihood tree, numbers at nodes represent percentage bootstrap values (only values above 50% are shown); **(B)** Bayesian tree, number at branch nodes refer to posterior probabilities.

retrieve other *Acsl4*-like sequences in other teleost species. *Acsl3* and *Acsl4* gene copies were also found in the cat shark and little skate (Figure 2).

The phylogenetic analysis also resolves a further inaccurate annotation in the western clawed frog. In the Ensembl database two ORFs are annotated as *Acsl4* genes (1-NP_001090679.1-ENSXETG00000033126 and 2-ENSXETG00000012429). After observing the localization of these two sequences in the phylogenetic tree, we find that one of the annotated *"Acsl4"* groups within the *Acsl3* clade, which is consistent with a synteny analysis. Therefore we consider that the western clawed frog presents one *Acsl4* gene (ENSXETG00000012429) and one inaccurate annotation of an *Acsl3* gene (ENSXETG00000033126) (Additional file 5).

In summary, the phylogenetic data indicates that the *Acsl1*, *Acsl2*, *Acsl5* and *Acsl6* and *Acsl3* and *Acsl4* have all duplicated before vertebrate radiation, with episodes of lineage specific expansion observed in studied invertebrate deuterostomes. In addition, teleost fish underwent specific duplications in *Acsl1* and *Acsl3*, and possibly *Acsl4* in zebrafish and in cave fish.

### Genome duplications contributed to the diversity of *Acsl* genes in vertebrates

Our phylogenetic analysis clearly indicates that despite the existence of several *Acsl* gene copies in the studied invertebrate deuterostome species, the expansion of the *Acsl1/Acsl2/Acsl5/Acsl6* and of the *Acsl3/Acsl4* clades took place in the vertebrate ancestor. Thus, we next analyzed the contribution of 2R and 3R in the generation of *Acsl* gene diversity. We started by examining the genomic location of each human *ACSL* gene and respective flanking gene families (Figure 3). Human *ACSL1, ACSL5 and ACSL6* localize respectively to Chr4q35, Ch10q25.2 and Ch5q31 (Figure 3A), regions which are part of the 2R NK-paralogon [24-26]. The analysis of the flanking gene families revealed that those which are multi-copy and whose duplication timing coincides with vertebrate emergence, typically have their members localizing to Hsa4, Hsa5, Hsa8, Hsa10 and/or Hsa2. For example, *TCF7L2* gene which flanks *ACSL5* has two other duplicates mapping to Hsa4 (*LEF1*) and Hsa5 (*TCF7*); *CASP3* which maps close to *ACSL1* has a paralogue, *CASP7*, mapping close to *ACSL5*; *PDLIM4* mapping downstream of *ACSL6* has two paralogues, *PDLIM1* and *PDLIM3*, localizing to

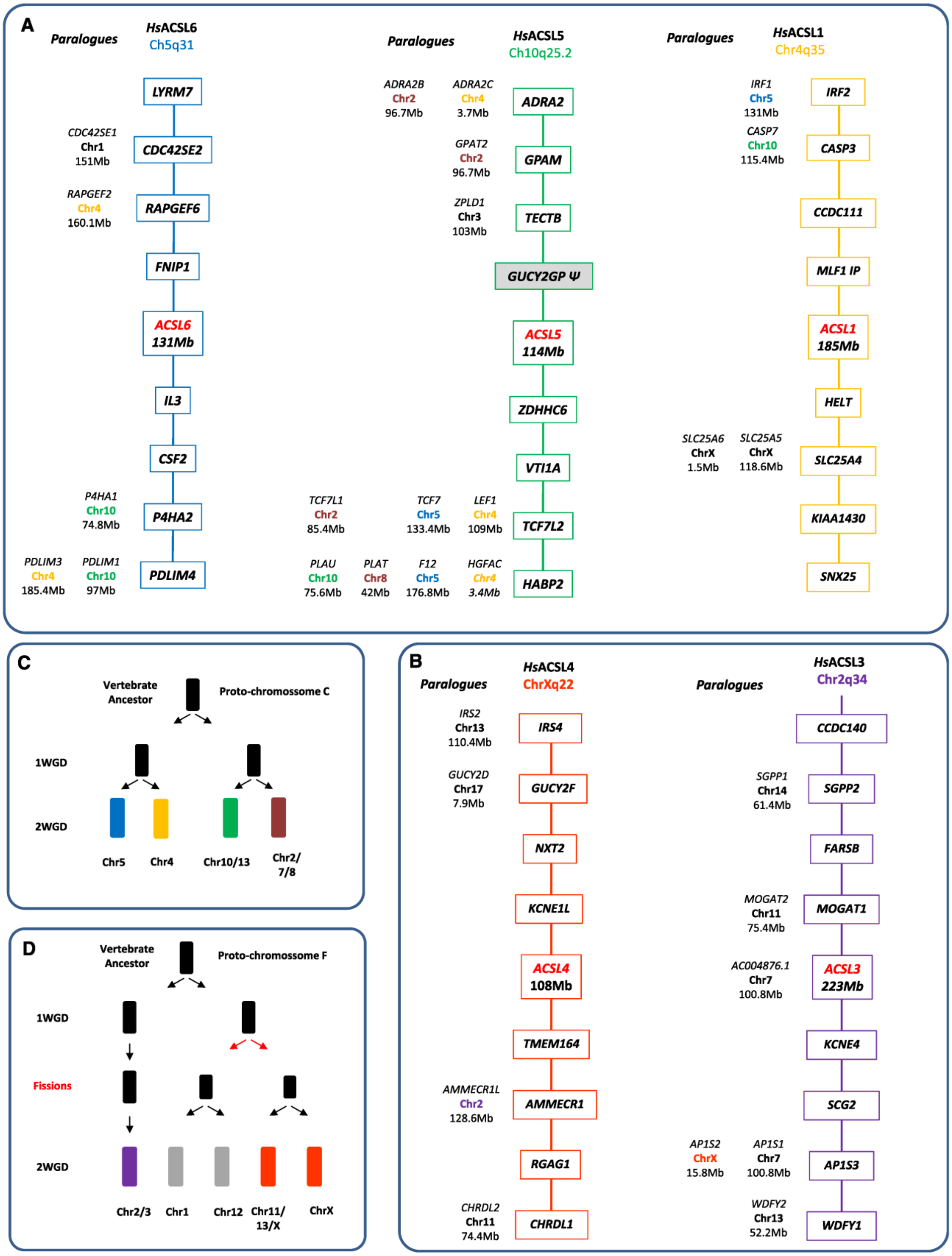


**Figure 3** (See legend on next page.)

(See figure on previous page.)
**Figure 3 Microsyteny analysis of the *Acsl* human *loci* and their mapping location in the ancestral vertebrate chromosomes. (A)** Location of the *Acsl1, Acsl5, Acsl6* and neighboring genes in the human genome and corresponding paralogues; **(B)** location of *Acsl3, Acsl4* and neighboring genes in the human genome and corresponding paralogues; **(C and D)** schematic representation of the duplication history of the ancestral vertebrate chromosomes C and F.

Hsa10 and Hsa4 respectively. Overall, the majority of genes flanking human *ACSL1*, *ACSL5* and *ACSL6* revealed conserved macrosynteny and therefore support the hypothesis that these regions are related, with the duplication timing coinciding with 2R. Furthermore, using the proposed vertebrate ancestral genome reconstruction [27], we find that the Hsa4, Hsa10 and Hsa5 belong to the same ancestral group, group C (Figure 3C). In summary, from a single ancestral chromosome C in the vertebrate ancestor, derived four chromosomes (C0, C1, C2 and C3) [27] as a result of 1R/2R. Each human *ACSL locus* maps to a distinct ancestral C chromosome: *ACSL1*-C1, *ACSL6*-C0 and *ACSL5*-C2 (Figure 3A and C). We would expect to find a fourth *ACSL* gene which should map to the ancestral chromosome C3 distributed in present day human genome at Hsa2/7/8 (see next section).

Regarding human *ACSL3* and *ACSL4* genes we find that they map to chromosomes Chr2q34 and ChrXq22 respectively. Neighboring gene families have paralogues in the following set of chromosomes HsaX, Hsa2, Hsa7, Hsa11, Hsa13, Hsa14 and Hsa17 (Figure 3B), with no apparent conserved macrosynteny. However, *ACSL3* and *ACSL4* map to chromosome regions derived from the 2R duplication of the proto-chromosome F, at F0 and F4 respectively (Figure 3D). Accordingly, after the first round of genome duplication one F proto-chromosome underwent an additional fission event, resulting in three proto-chromosomes. Two of these proto-chromosomes underwent the second WGD, giving rise to four ancestral chromosomes (F1, F2, F3 and F4) and the third chromosome gave rise to the F0. The gene families flanking *ACSL3*/*ACSL4* have in most cases duplicates in regions assigned to F chromosomes [27], thus providing strong support to the hypothesis that both were 2R generated.

Extra gene copies of *Acsl1*, *Acsl3*, and *Acsl4* (in zebrafish) were found in our survey. The analysis of the *loci* of *Acsl1*, *Acsl3* and *Acsl4* (Figure 4) in stickleback and zebrafish provides solid support to the conclusion that 3R contributed for the increase of the *Acsl* gene set in teleosts. We find that 3R specific duplicates can be observed outflanking each pair of *Acsl* duplicates (*Casp3*, *Ephb1* and *Stag2*) (Figure 4).

### *Acsl2* is a potential 2R paralogue gone missing in tetrapods

The phylogenetic analysis showed the presence of a previously unidentified *Acsl* gene, *Acsl2*, paralogous to *Acsl1*, *Acsl5* and *Acsl6*. To enlighten the evolutionary origin of the *Acsl2* gene we analyzed the genomic *locus* of this novel gene in teleost species (Figure 5). We show that the *Acsl2* gene is confined to a largely conserved *locus* in teleost fish. A large set of neighboring gene families have their human orthologues mapping to Hsa8. The following genes *EGR3*, *BIN3*, *RHOBTB2*, *BMP1*, *PEBP4*, *STC1*, and *IDO2* are close together at Hsa8 and constitute a conserved block, with *UBL4b* and *PTCHD2* localizing in Hsa1 (Figure 5; data not shown). Most importantly, the conserved syntenic block in Hsa8 maps back to the ancestral chromosome C3 which corresponds to the expected localization of a fourth copy of the *ACSL* gene after 2R, absent in the human genome. Further, gene families which are multicopy have their paralogues localizing to Hsa10/5/4 as expected. These finds together with the phylogenetic analysis suggest that the *Acsl2* gene corresponds to a retained paralogue conserved in teleosts and lost in the tetrapod lineage.

### Gene expression data indicates functional partitioning and diversification

Given that the teleosts have additional *Acsl* gene copies, we proceeded to analyze the gene expression of *Acsl* genes in zebrafish and performed a comparative analysis with the human *ACSL* gene repertoire. In zebrafish a high *Acsl1a* mRNA content is observed in all analyzed tissues with the exception of the eye (Figure 6A), while *Acsl1b* is only marginally expressed in testis, ovary, kidney and heart (Figure 6A). The human *ACSL1* has a similar expression pattern with high expression in brain, heart, spleen, kidney, ovary and testis and medium to low expression in liver, lung and eye (Figure 5C). These findings are in agreement with previous studies in *Rattus norvegicus* in which it was found that *Acsl1* is highly expressed in major energy metabolizing tissues namely heart, liver and adipose tissues [28]. Regarding ACSL5, in human we find that this enzyme is highly expressed in all tissues here analyzed with the exception of the ovary (Figure 6C). When observing the data obtained for zebrafish we find a distinct expression pattern. High *Acsl5* mRNA transcription is observed in the testis, ovary, kidney, gut and liver, while spleen, gill, heart, eye and brain have a low/absent gene transcription (Figure 6A).

Concerning ACSL6, in opposition to ACSL1 and ACSL5; this enzyme presents a fairly restricted expression pattern in human being highly expressed in testis, ovary and brain.

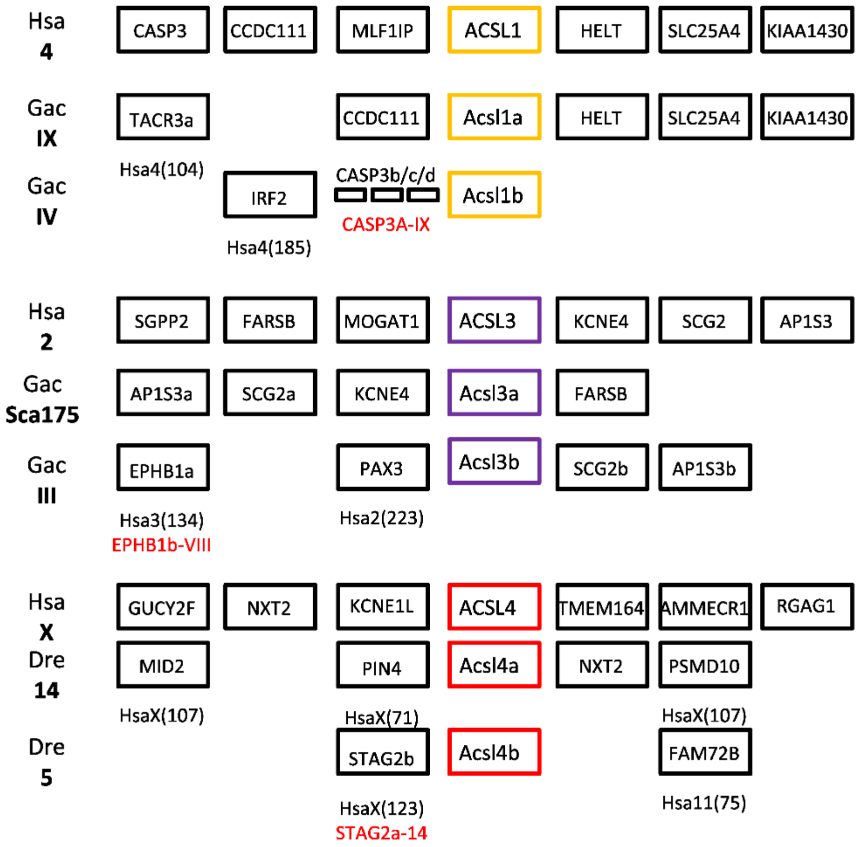


**Figure 4** **Conserved synteny of *Acsl1*, *Acsl3* and *Acsl4* between human and teleost (*G. aculeatus* and *D. rerio*) indicate the contribution of 3R in the generation of extra gene copies.**

In zebrafish this restricted expression pattern is also observed, with *Acsl6* being found essentially in brain (Figure 6A). The teleost *Acsl2* transcription is high in testis, ovary, gill, heart, eye, and brain.

The expression pattern of *Acsl3a* and *Acsl3b* in zebrafish, reveals that *Acsl3a* is preferably expressed in ovary, gill and brain; nevertheless this gene is also expressed at lower levels in all other tissues (Figure 6B). *Acsl3b* is expressed in all tissues with comparatively higher levels to *Acsl3a,* with the exception of eye. The expression pattern of the *Acsl4a* and *Acsl4b* is highly similar being highly expressed in all tissues here analyzed with a slight decrease in gut and liver and low expression in eye (Figure 6B). When observing the expression pattern of *ACSL4* in human we observe that this gene is also highly expressed in all tissues with the exception of gut and eye (Figure 6D).

## Discussion

ACSL are key enzymes involved in the initial steps of FA metabolism. These enzymes preferentially activate FA with C12-C20 (the most abundant in the human diet), which may then intervene in a variety of metabolic pathways. Although the *ACSL* family has been the focus of various studies, their evolutionary history has not been investigated before. Here we combine extensive database search with phylogenetics and comparative genomics to provide a reliable depiction of the evolution of *Acsl* in vertebrates (Figure 7). Initial analyses revealed several inaccurate gene annotations. After an exhaustive analysis we were able to clarify several of these and perceive that the diversity of *Acsl* genes is broader than anticipated. The gene repertoire varies significantly between mammals (5), teleosts (7/8), and the invertebrate studied species (3/4). Through phylogenetics we were able to reconstruct the *Acsl* gene duplication timings in relation to the divergence of major vertebrate groups. The five mammalian *Acsl* genes have been organized into two separate groups: *Acsl1, Acsl5, Acsl6* and *Acsl3, Acsl4,* on the basis of sequence homology and gene organization [10,12,18]. We now propose that this division is evolutionarily significant and dates back to at least deuterostome ancestry since clear co-orthologues of both gene groups exist in hemichordates and cephalochordates. Although, various *Acsl1/2/5/6* genes were found in amphioxus and the acorn worm, these represent

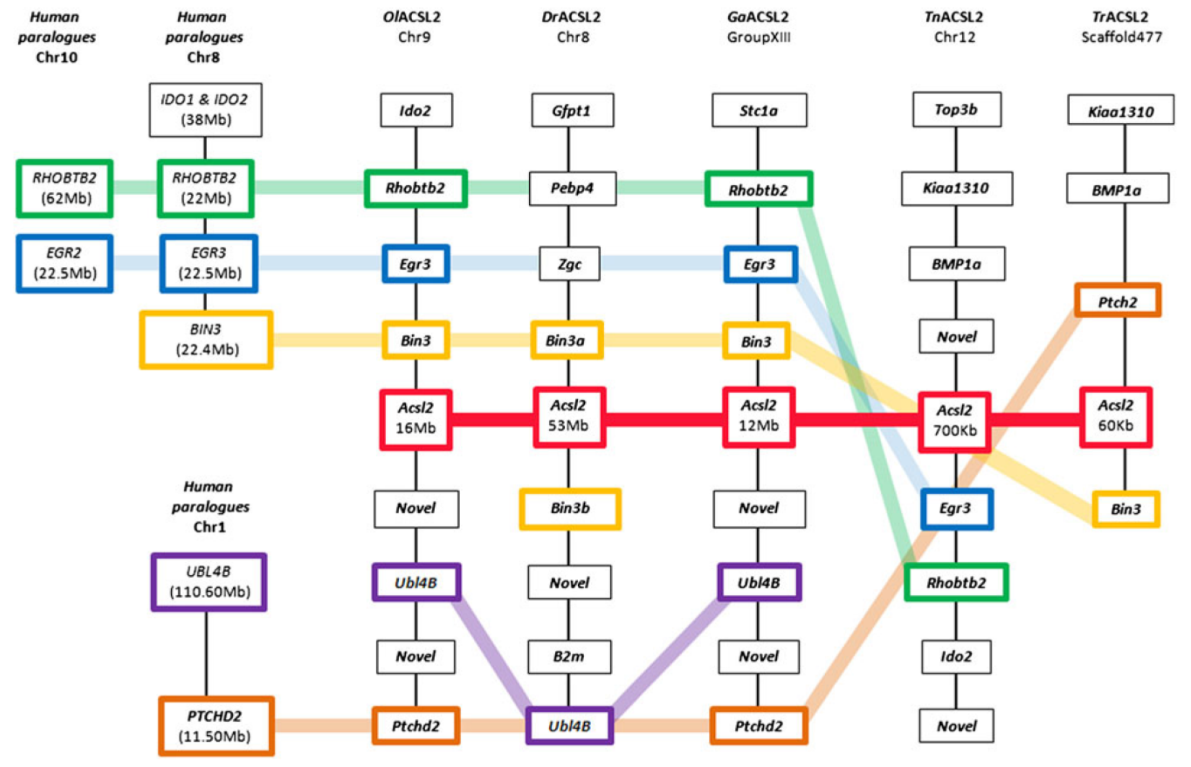


**Figure 5** Synteny map of *Acsl2 loci* in teleost fish, and the location of human orthologues.

independent lineage duplications. The exact duplication timing of a proto-*Acsl* gene to originate the ancestor of *Acsl1/2/5/6* and *Acsl3/4* is at present unknown but probably dates as far back as the origin of the Bilateria (not shown; LFCC unpublished results). In the agnathan lamprey we were only able to retrieve one complete *Acsl* sequence, although several partial sequences were also evident. Thus, at this stage a final conclusion concerning the full repertoire of *Acsl* genes in lamprey is premature. The findings that the *Acsl* gene family expanded significantly in the time window coincident with the emergence of vertebrates lead us to test the contribution of genome duplications. Using the proposed ancestral vertebrate genome reconstruction [27] we were able to trace back the duplication history of *Acsl* genes in the gnathostome ancestor. We find that human *Acsl1, Acsl5 and Acsl6*, map to chromosomes C1, C2 and C0 respectively which originated from duplication of the ancestral proto-chromosome C (Figure 7). This observation is also supported by the duplication history analysis of the flanking gene families. In teleosts we found extra gene copies within the *Acsl1/5/6* clade. These were partially explained by the contribution of the teleost specific genome duplication (3R) (*Acsl1a* and *Acsl1b*), but not entirely. A novel gene with no clear orthologues in tetrapods was found in the analyzed teleost species, the spotted gar, coelacanth and possibly lamprey. The phylogenetic analysis clearly indicated that this represents a distinct gene lineage which we name *Acsl2*. To enlighten the origin of *Acsl2*, we investigated the genomic *locus* in teleosts and cross-compared it with the human genome. We find that the most parsimonious explanation for the retrieved data is that *Acsl2* is a 2R paralogue retained in teleosts and lost in tetrapods, similar to what was found in a distinct ACS gene family [21]. Thus, the novel uncharacterized *Acsl2* gene corresponds to a quadruplicate *Acsl* paraloguous to *Acsl1, Acsl5* and *Acsl6.*

Similarly, we found also that the human orthologues of *Acsl3* and *Acsl4* map to chromosome regions related by duplication. Both genes map to genomic regions remnants of 2R resulting from the duplication of the ancestral proto-chromosome F. The most plausible explanation for the unequal distribution of *Acsl* gene copies within F0, F1, F2, F3 and F4, is a chromosome fission event occurred after 1R of WGD which resulted in the splitting of the genetic information into two distinct chromosomes, which then underwent the second genome duplication (2R) and originated F1, F2, F3 and F4 (Figure 7). The detailed comparative genomic and phylogenetic analysis again highlighted the contribution

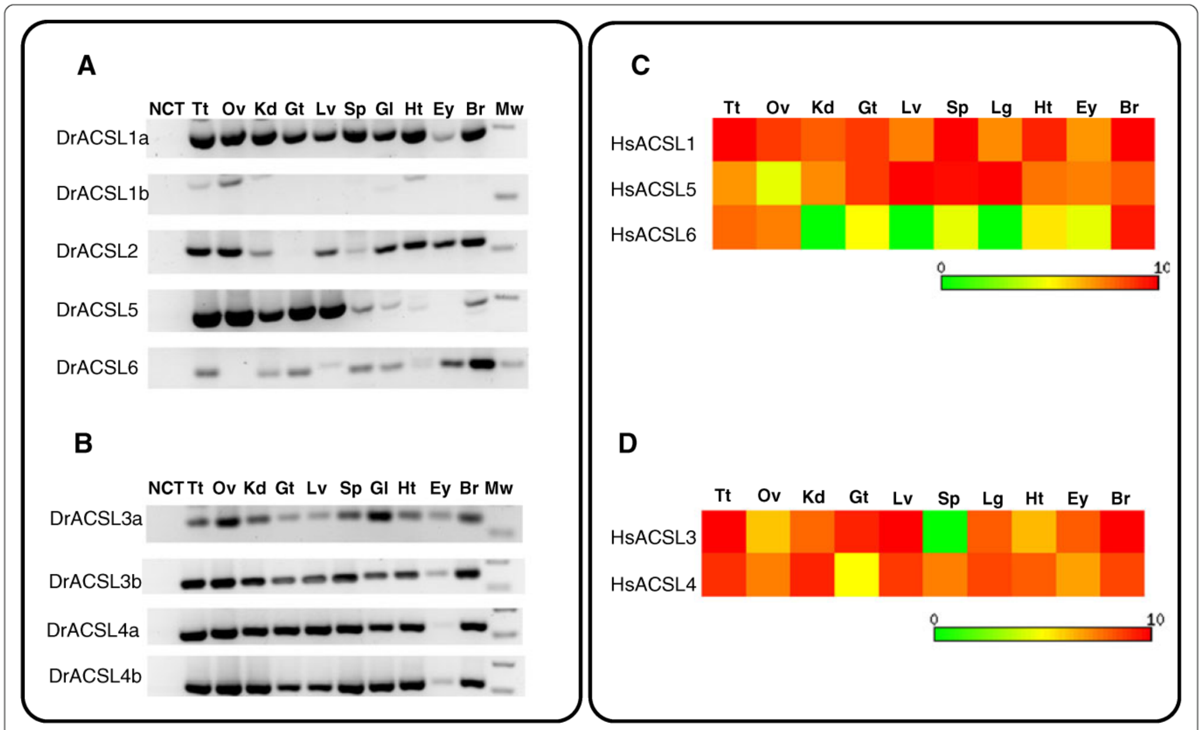


**Figure 6 Comparative expression analysis of *Acsl* genes in human and zebrafish tissues. (A and B)** Tissue expression profile of zebrafish *Acsl* genes; **(C and D)** Tissue expression heatmap of Human *Acsl* genes; NCT negative template control, Tt: testis, Ov: ovary, Kd: kidney, Gt: gut, Lv: Liver, Sp: spleen, Gl: gill, Ht: heart, Ey: eye; Br: brain, Lg: lung, Mw: molecular weight marker.

of 3R to the gene number increment observed in teleosts (Figure 7).

It has been previously suggested that major vertebrate innovations occurred after genome duplication events [5]. Whole genome duplications lead to the expansion of gene numbers, facilitating gene diversification, subfunctionalization along with the rise of novel functions and gene loss. These ultimately enable evolutionary radiation. FA composition and metabolism is known to be different in some vertebrate groups, for example in teleosts [20]. Similar to our findings in *Acsl*, recent studies have also revealed that various gene families involved in lipid metabolic pathways have evolved distinct gene repertoires in vertebrate lineages, including fish, with clear functional and regulatory impacts [20,21,29,30]. The retention of such a larger *Acsl* gene set after 2R/3R, with simultaneous processes of differential loss, could be indicative that novel *Acsl* gene functions have emerged in vertebrate ancestry. In effect, the variety of *ACSL*s in mammals is apparently associated with distinct substrate preferences. ACSL1 uses FA with C16 to C18 both saturated and mono unsaturated, ACSL3 displays a high activity with C12:0 (laurate), C14:0 (myristatate), C20:4 (arachidonate) and C20:5 (eicosapentaenoic acid) [31]. In contrast, ACSL4 presents a clear preference for polyunsaturated FA with C2O:4 and C20:5 [10,32]. ACSL5 shows substrate specificity similar to ACSL1, favorably utilizing palmitic (C16:0), palmitoleic (C16:1), oleic (C18:1) and linoleic (C18:2) acids [33]. Finally, ACSL6 preferentially uses FAs with C16 to C20 saturated and polyunsaturated, although alternative splicing generates isoforms with distinct substrate specificities [10,34]. We propose that teleost orthologues have probably retained similar FA substrate preferences with respect to *Acsl*s. However, novel *Acsl* family members were also discovered in this study. Thus, we performed a comparative tissue transcription analysis between zebrafish and human. We selected zebrafish as a model, given that this species has been previously suggested as a model organism for the study of lipid metabolism [22], and coincidentally this species also presents the largest set of *Acsl* genes. The comparison of the expression data between human and zebrafish revealed a similar expression profile, with the exception of *Acsl2* and *Acsl5,* and the fish specific duplicates (*Acsl3* and *Acsl4*). We observe that the zebrafish *Acsl5* is expressed in fewer tissues when compared to the human orthologue, with the teleost *Acsl2* apparently compensating the lack of *Acsl5* transcription in various tissues. This setting suggests that in teleost fish functions are shared between *Acsl2* and *Acsl5*. Regarding the 3R duplicated

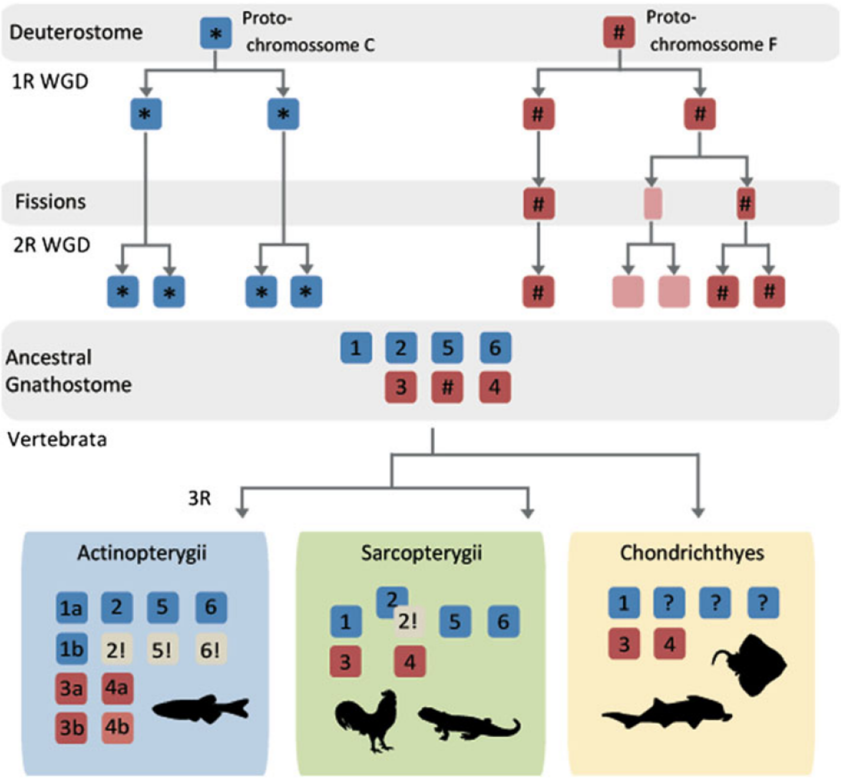


**Figure 7 Proposed evolutionary history and duplication timing of the *Acsl* gene family in vertebrates.** Questions marks indicates unknown data, and exclamation signals gene loss.

members *Acsl3a* and *Acsl3b*, and *Acsl4a* and *Acsl4b* we find similar tissue expression distributions in opposition to *Acsl1a* and *Acsl1b*. In the latter *Acsl1b* is co-expressed with *Acsl1a* in a small set of specific tissues (testis, ovary and heart). We hypothesize that *Acsl1b* plays a specific role in these tissues, distinct from the role played by *Acsl1a*, hinting towards a sub-functionalization after duplication. Although, we have not tested the FA specificity of the novel *Acsl* repertoire described in this work, we cannot ignore that the retention of a larger *Acsl* gene number in teleosts could also be related with the specific acquisition of novel substrate preferences, which future studies should address.

## Conclusion

In summary, we demonstrate the importance of genome duplications, 2R and 3R, in the generation of the Acsl diversity in vertebrate species.

## Methods

### Database mining and identification of Acsl sequences

ACSL family members were identified in the Ensembl, GenBank and JGI (Joint genome institute) databases through Blastp searches using as reference annotated human ACSL sequences. In order to include all major vertebrate lineages we analysed eutherian metatherian and prototherian mammals: *Homo sapiens* (human), *Mus musculus* (mouse); *Monodelphis domestica* (opossum); *Ornithorhynchus anatinus* (platypus); birds: *Gallus gallus* (chicken); reptiles: *Anolis carolinensis* (anole lizard); amphibians: *Xenopus tropicalis* (western clawed frog); Latimeria chalumnae (Coelacanth); *Lepisosteus oculatus* (spotted gar); teleosts: *Danio rerio* (zebrafish), *Astyanax mexicanus* (blind cave fish), *Takifugu rubripes* (pufferfish), *Tetraodon nigroviridis* (green spotted puffer) *Oryzias latipes* (medaka) and *Gasterosteus aculeatus* (stickleback); chondrichthyans: *Leucoraja erinacea* (little skate), *Scyliorhinus canicula* (small-spotted catshark) and *Callorhinchus milii* elephant shark,and jawless fish hyperoartia: *Petromyzon marinus* (sea lamprey). Sequences searches were also made in an invertebrate chordate *Branchiostoma floridae* (amphioxus) and the hemichordate *Saccoglossus kowalevskii* (acorn worm).

### Sequence alignment and phylogenetic analysis

All ACSL amino acid sequences retrieved in the database mining were initially aligned in MAFFT alignment software using default parameters [35] and manually curated with the exclusion of regions of uncertain homology, gaps, and of partial sequences. Revised sequence alignments were then submitted to Protest online server version 2.4 [36] available at http://darwin.uvigo.es/software/prottest_server.html, in order to select the appropriate protein evolution model according to our dataset. Here we found that *ACSL3* and *ACSL4* group follows a JTT + I + G model, while the LG + I + G + F model suits best the *ACSL1*, *ACSL2 ACSL5* and *ACSL6* group. Phylogenetic analyses were performed and a Maximum likelihood (PhyML) tree with 1000 bootstrap replicates was constructed using the online platform of the PhyML 3.0 avaliable at http://www.atgc-montpellier.fr/phyml/. Bayesian inference of phylogeney was performed with MrBayes version 3.2.2 [37] on CIPRES Science Gateway [38]. Analysis for both Acsl3/4 and Acsl1/2/5/6 amino acid sequences were performed under a mixed substitution model, with two parallel runs with 1 million generations, each with four chains one cooled and 3 heated. Trees were sampled every 100 generations; final consensus tree was calculated with the fifty percent majority rule and from the remaining trees after a 0.25 burin.

### Comparative genomics

All *ACSL* genes were mapped into the human chromosomes, the location of each gene and the neighboring genes were collected from Ensembl and GenBank databases. *ACSL loci* in human were used as a model for comparison. The Ensembl paralogue and orthologue prediction tools were used to infer duplication history patterns of flanking ACSL genes. For some flanking gene families we run phylogenetics to confirm relationships with the methodology described above (not shown).

### Gene transcription analysis

Adult wild-type zebrafish obtained from our own breeding stock were used for gene expression analysis. Animals were anesthetized and killed by cervical transection in accordance with the Portuguese Animals and Welfare Law (Decreto-Lei nº 197/96) approved by the Portuguese Parliament in 1996. Institutional animal approval by CIIMAR/UP and DGV (Ministry of Agriculture) was granted for this study. After collection all tissues were preserved in RNAlater and stored at −20°C. Total RNA was isolated using an Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, UK) according to the manufactures recommendations, including the on-column treatment of isolated RNA with RNase-free DNase I. RNA concentration was calculated using Qubit fluorometer instrument (Invitrogen, Carlsbad CA), integrity confirmed by electrophoresis and the RNA stored at −80°C until further use. The cDNA was synthesized from 250 ng of total RNA with the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufactures protocol. Forward and reverse primers were designed using sequences available in Ensembl with Primer3 software [39]. All of primer sets match exon sequence and flank an intron consequently avoiding genomic DNA amplification. Primers sets were created for the following genes *ACSL1a*; (Forward-5′ CAGGATGGGCAAAGAATAGAG 3′, Reverse-5′ TTT CAGTGTTGGTGTGAGGAG 3′, annealing at 55°C) *ACSL1b;* (Forward-5′ GCACAGCGAGATGTTCAC 3′, Reverse-5′ AAGTCCAATCCAAATGTCAGG 3′, annealing at 54°C) *ASCL2* gene; (Forward-5′ GTAGTTCCAGATCCA GAAGTGTTC 3′ reverse-5′ CGCCGTCATGTCCTCCAG 3′, annealing at 56°C) ACSL5; (Forward-5′ CGCAGAGAA ACTGGGATTGAAAGG 3′, Reverse-5′ TGGCTTTGAGT GTTGGAGTGAGG 3′, annealing at 58°C) and *ASCL6* (Forward-5′ CCTCGTGGGCTCAGAAGAAAG 3′ Reverse-5′CGCACCATGTCCTCCAGAATA 3′, annealing at 58°C). PCR was performed using 2 μl of zebrafish cDNA and Phusion® Flash high-fidelity Master Mix (FINNZYMES). PCR parameters were as follows: initial denaturation at 98°C for 10 s, followed by 35 cycles of denaturation at 98°C for 1 s, annealing for 5 s and elongation at 72°C for 10s and a final step of elongation at 72°C for 1 min. PCR products were then loaded onto 2% agarose gel stained with GelRed and run in TBE buffer at 80 V. *In silico* expression analysis, for *ACSL* gene in Human, was performed using ESTs available from Unigene [40] as count per million transcripts, all values are displayed as Log2 transcripts per million. Heat map was created using the collected EST data and matrix2png web interface v1.2 [41].

## Additional files

**Additional file 1: tBlastn search of ACSLlike sequences in Transcriptomic contigs.**

**Additional file 2: Partial *Acsl* gene annotations in green spotted pufferfish.**

**Additional file 3: NCBI accession numbers and Ensembl gene ID.**

**Additional file 4: Synteny maps of Zebrafish ACSL 3R duplicates.**

**Additional file 5: Xenopus tropicalis *Acsl4* and *Acsl3* corresponding location in human.**

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

The original idea for this study was conceived by LFCC. LFCC and MLM performed all the experimental analysis; MMS, MARH, and IC participated in the discussion regarding lipid metabolism and physiology. The manuscript was written by LFCC and MLM, and edited by all other co-authors.
All authors have read and approved the final manuscript.

#### Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia (Portugal), through grants PTDC/MAR/68106/2006 and PTDC/MAR/68885/2006, the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT – Fundação para a Ciência e a Tecnologia, under the project "PEst-C/MAR/LA0015/2013, and SFRH/BD/84238/2012 awarded to M Lopes-Marques. We thank three anonymous reviewers for their insightfull comments.

**Author details**
[1]CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, CIMAR Associate Laboratory, UPorto, University of Porto, Porto, Portugal. [2]ICBAS (Instituto de Ciências Biomédicas Abel Salazar), University of Porto, Porto, Portugal. [3]Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal.

Received: 19 December 2012 Accepted: 4 December 2013
Published: 12 December 2013

**References**

1. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, *et al*: **The amphioxus genome and the evolution of the chordate karyotype.** *Nature* 2008, **453**(7198):1064–1071.
2. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, *et al*: **Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.** *Nature* 2004, **431**(7011):946–957.
3. Moghadam HK, Ferguson MM, Danzmann RG: **Whole genome duplication: challenges and considerations associated with sequence orthology assignment in Salmoninae.** *J Fish Biol* 2011, **79**(3):561–574.
4. Crow KD, Smith CD, Cheng J-F, Wagner GP, Amemiya CT: **An independent genome duplication inferred from Hox paralogs in the American paddlefish-a representative basal ray-finned fish and important comparative reference.** *Genome Biol Evol* 2012.
5. Shimeld SM, Holland PWH: **Vertebrate innovations.** *Proc Natl Acad Sci* 2000, **97**(9):4449–4452.
6. Postlethwait JH: **The zebrafish genome in context: ohnologs gone missing.** *J Exp Zool B Mol Dev Evol* 2007, **308B**(5):563–577.
7. Mulley JF, Holland PWH: **Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths.** *Mol Biol Evol* 2010, **27**(10):2386–2391.
8. Kuraku S, Kuratani S: *Genome-Wide Detection of Gene Extinction in Early Mammalian Evolution*. Evolution: Genome Biology and; 2011.
9. Widmark J, Sundström G, Ocampo Daza D, Larhammar D: **Differential evolution of voltage-gated sodium channels in tetrapods and teleost fishes.** *Mol Biol Evol* 2011, **28**(1):859–871.
10. Soupene E, Kuypers FA: **Mammalian long-chain acyl-CoA synthetases.** *Exp Biol Med* 2008, **233**(5):507–521.
11. Watkins PA, Ellis JM: **Peroxisomal acyl-CoA synthetases.** *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 2012, **1822**(9):1420–1411.
12. Watkins PA, Maiguel D, Jia Z, Pevsner J: **Evidence for 26 distinct acyl-coenzyme A synthetase genes in the human genome.** *J Lipid Res* 2007, **48**(12):2736–2750.
13. Li LO, Klett EL, Coleman RA: **Acyl-CoA synthesis, lipid metabolism and lipotoxicity.** *Biochimica et Biophysica Acta (BBA) - Mol Cell Biol Lipids* 2009, **1801**(3):246–251.
14. Morino K, Petersen KF, Shulman GI: **Molecular mechanisms of insulin resistance in humans and their potential links with mitochondrial dysfunction.** *Diabetes* 2006, **55**(Supplement 2):S9–S15.
15. Longo I, Frints SGM, Fryns JP, Meloni I, Pescucci C, Ariani F, Borghgraef M, Raynaud M, Marynen P, Schwartz C, *et al*: **A third MRX family (MRX68) is the result of mutation in the long chain fatty acid-CoA ligase 4 (FACL4) gene: proposal of a rapid enzymatic assay for screening mentally retarded patients.** *J Med Genet* 2003, **40**(1):11–17.
16. Meloni I, Muscettola M, Raynaud M, Longo I, Bruttini M, Moizard M-P, Gomot M, Chelly J, des Portes V, Fryns J-P, *et al*: **FACL4, encoding fatty acid-CoA ligase 4, is mutated in nonspecific X-linked mental retardation.** *Nat Genet* 2002, **30**(4):436–440.
17. Cao Y, Pearman AT, Zimmerman GA, McIntyre TM, Prescott SM: **Intracellular unesterified arachidonic acid signals apoptosis.** *Proc Natl Acad Sci* 2000, **97**(21):11280–11285.
18. Mashek DG, Bornfeldt KE, Coleman RA, Berger J, Bernlohr DA, Black P, DiRusso CC, Farber SA, Guo W, Hashimoto N, *et al*: **Revised nomenclature for the mammalian long-chain acyl-CoA synthetase gene family.** *J Lipid Res* 2004, **45**(10):1958–1961.
19. Grove TJ, Sidell BD: **Fatty acyl CoA synthetase from Antarctic notothenioid fishes may influence substrate specificity of fat oxidation.** *Comp Biochem Physiol B Biochem Mol Biol* 2004, **139**(1):53–63.
20. Morais S, Monroig O, Zheng X, Leaver M, Tocher D: **Highly unsaturated fatty acid synthesis in Atlantic salmon: characterization of ELOVL5- and ELOVL2-like elongases.** *Marine Biotechnol* 2009, **11**(5):627–639.
21. Castro LFC, Lopes-Marques M, Wilson JM, Rocha E, Reis-Henriques MA, Santos MM, Cunha I: **A novel acetyl-CoA synthetase short-chain subfamily member 1 (Acss1) gene indicates a dynamic history of paralogue retention and loss in vertebrates.** *Gene* 2012(0).
22. Flynn EJ, Trent CM, Rawls JF: **Ontogeny and nutritional control of adipogenesis in zebrafish (Danio rerio).** *J Lipid Res* 2009, **50**(8):1641–1652.
23. Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH: **Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted Gar, an outgroup for the teleost genome duplication.** *Genetics* 2011, **188**(4):799–808.
24. Pollard SL, Holland PWH: **Evidence for 14 homeobox gene clusters in human genome ancestry.** *Curr Biol* 2000, **10**(17):1059–1062.
25. Castro LFC, Holland PWH: **Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes.** *Evol Dev* 2003, **5**(5):459–465.
26. Luke GN, Castro LFC, McLay K, Bird C, Coulson A, Holland PWH: **Dispersal of NK homeobox gene clusters in amphioxus and humans.** *Proc Natl Acad Sci* 2003, **100**(9):5292–5295.
27. Nakatani Y, Takeda H, Kohara Y, Morishita S: **Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates.** *Genome Res* 2007, **17**(9):1254–1265.
28. Mashek DG, Li LO, Coleman RA: **Rat long-chain acyl-CoA synthetase mRNA, protein, and activity vary in tissue distribution and in response to diet.** *J Lipid Res* 2006, **47**(9):2004–2010.
29. Castro LFC, Monroig Ó, Leaver MJ, Wilson J, Cunha I, Tocher DR: **Functional Desaturase Fads1 (Δ5) and Fads2 (Δ6) Orthologues Evolved before the Origin of Jawed Vertebrates.** *PLoS ONE* 2012, **7**(2):e31950.
30. Castro LFC, Wilson J, Goncalves O, Galante-Oliveira S, Rocha E, Cunha I: **The evolutionary history of the stearoyl-CoA desaturase gene family in vertebrates.** *BMC Evol Biol* 2011, **11**(1):132.
31. Fujino T, Kang M-J, Suzuki H, Iijima H, Yamamoto T: **Molecular characterization and expression of Rat acyl-CoA synthetase 3.** *J Biol Chem* 1996, **271**(28):16748–16752.
32. Kang M-J, Fujino T, Sasano H, Minekura H, Yabuki N, Nagura H, Iijima H, Yamamoto TT: **A novel arachidonate-preferring acyl-CoA synthetase is present in steroidogenic cells of the rat adrenal, ovary, and testis.** *Proc Natl Acad Sci* 1997, **94**(7):2880–2884.
33. Oikawa E, Iijima H, Suzuki T, Sasano H, Sato H, Kamataki A, Nagura H, Kang M-J, Fujino T, Suzuki H, *et al*: **A novel acyl-CoA synthetase, ACS5, expressed in intestinal epithelial cells and proliferating preadipocytes.** *J Biochem* 1998, **124**(3):679–685.
34. Soupene E, Dinh N, Siliakus M, Kuypers F: **Activity of the acyl-CoA synthetase ACSL6 isoforms: role of the fatty acid Gate-domains.** *BMC Biochem* 2010, **11**(1):18.
35. Katoh K, Toh H: **Parallelization of the MAFFT multiple sequence alignment program.** *Bioinformatics* 2010, **26**(15):1899–1900.
36. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2009, **21**(9):2104–2105.
37. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19**(12):1572–1574.
38. Miller MA, Pfeiffer W, Schwartz T: **Creating the CIPRES Science Gateway for inference of large phylogenetic trees.** *Proc Gateway Comput Environ Workshop (GCE)* 2010:1–8.
39. Rozen S, Skaletsky H: **Primer3 On the WWW for general users and for biologist programmers.** 1999, **132:**365–386.
40. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, *et al*: **Database resources of the national center for biotechnology information.** *Nucleic Acids Res* 2010, **38**(suppl 1):D5–D16.
41. Pavlidis P, Noble WS: **Matrix2png: a utility for visualizing matrix data.** *Bioinformatics* 2003, **19**(2):295–296.

doi:10.1186/1471-2148-13-271
**Cite this article as:** Lopes-Marques *et al.*: Diversity and history of the long-chain acyl-CoA synthetase (*Acsl*) gene family in vertebrates. *BMC Evolutionary Biology* 2013 **13**:271.

## SUPPLEMENTARY MATERIAL

Additional file 1

### tBlastn search of ACSLlike sequences in Transcriptomic contigs

tBlastn searches were performed on Transcriptome Contigs in elephant shark, little skate, small spotted catshark (http://skatebase.org) using as reference human ACSL protein sequences, contigs containing hits were retrieved. Ensembl database was used to retrieve annotated ACSL genes in Lamprey and Coelacanth while Pré-Ensembl database was used for Spotted Gar.
Sequence sorting and profiling was done using phylogenetic analysis with partial sequences and Blastp in NCBI

| | Acsl1/Acsl2/Acsl5/ Acsl6 | Acsl3/Acsl4 |
|---|---|---|
| Elephant shark | ctg60368 (f)ACSL1<br>ctg48116 (p) ACSL6? | ctg17226 (p)ACSL3?<br>ctg79717 (p)ACSL3?<br>ctg24956(p)ACSL4?<br>ctg26023 (p)ACSL4?<br>ctg18935(p)ACSL4?<br>ctg34934 (p)ACSL4? |
| Little Skate | ctg11347 (p) ACSL1?<br>ctg17167 (p) ACSL1?<br>ctg82046 (p) ACSL2?<br>ctg34321 (p) ACSL5?<br>ctg43339 (p) ACSL5?<br>ctg31550 (p) ACSL5? | ctg12475 (p)ACSL3?<br>ctg14318 (f) ACSL4 |
| Cat shark | ctg18625 (p) ACSL1?<br>ctg94905 (p) ACSL1?<br>ctg63419 (p)ACSL2?<br>ctg81762 (p)ACSL6? | ctg67516 ACSL3?<br>ctg66619ACSL4? |
| Lamprey | ENSPMAG00000008135 (f) ACSL1<br>ENSPMAG00000004625 (p) ACSL?<br>ENSPMAG00000005099 (p) ACSL? | ENSPMAG00000005133 (p) ACSL3 |
| Coelacanth | ENSLACG00000008655 (f) ACSL1<br>ENSLACG00000012732 (f) ACSL2<br>ENSLACG00000010845 (f) ACSL5<br>ENSLACG00000005834 (f) ACSL6 | ENSLACG00000003984 (f) ACSL3<br>ENSLACG00000002977 (f) ACSL4 |
| Spotted Gar | ENSDARP00000053749_1 (f) ACSL2<br>ENSDARP00000123999_1 (f) ACSL5<br>ENSDARP00000039916_1 (f) ACSL6 | ENSDARP00000042883_1 (f) ACSL3<br>ENSDARP00000010373_1 (f) ACSL4 |

**Notes**: (f) Indicates full sequences (p) Indicates partial sequences

Full sequences were further used in phylogenetic analysis in figure 1 and 2.

**Elephant shark**

```
>Cmictg60368ACSL1

METHELINHLRVPELWELRQYIRSLPTYTLMGIGAIAALVTYWYATRTKPPKPACDLSMQSIEVASGERA
RRSALLNSDDLLTRYYEDITTLYEVLKRGLRVSRNGPCLGSRKPNQPYEWMSYTEVSDKAEFLGSGLINQ
GCKATNDQFIGIFAQNRPEWVIVEQACYAYSMVVVPLYDTLGSEAISFILKQAEIAFVFCDTSIKAHSLL
AGVERGQTPLLKVIIVMDPIGNDLQERGRNCGVKIVHFKEVEDEGREQKRPVVPPKPSDLAVVCFTSGTT
GNPKGAMLTHGNIVSNLSAFLKVTEKVGVPTPADSLISFLPLGHMFERIVEAVVLCHGARIGFFQGDIRL
LMDDLKALKPTVLPVVPRLLNRMFDKVHNQANGLIKRQVLEFAAWRKEAQLQNGIVCRDTIWDKLVFHKV
QDNLGGRVKLMVTGAAPVSETVLTFLRAAIGCQFYEGYGQTECTAGCTMSIPGDWTAGHVGAPMPCNIIK
LIDVEDMDYFAVKGEGEVCVKGPNVFKGYLKDPEKTAEALDADGWLRTGDIGKWLPNGTLKIIDRKKHIF
KLAQGEYIAPEKIENVYVRSTPVAQVYVYGESLQACLVGIVVPDPETFSEWAQKTGVRGAYEELCKNQKV
KQAVLEDMVALGKESGLKSFEQVKDIVLHPEMLTVQSGLLTPTFKAKRAELRKYFRSQLDELYANIKM

>Cmictg17226ACSL3

LGQQSKSNIAIFCETRAEWMITAQACFMYNFPLVTLYATLGGNAIAHGLNETQITHIITSKKLLQTKLKD
ILLNVPKLQHIIIVDDKPTAWADYPRGIMVHNMTAVEDLGSKPENLHKQYTKPTTEDIAVIMYTSGSTGI
PKGVMISHGNLIAGIAGMSARVSGLGPKDTYVGYLPLAHVLELSAETLCLACGCRIGYSSPQTLADQSAK
IKKGSKGDTTVLRPTLLAAVPEIMDRIYKNIVNKVEEMNSIKRTLFVLAYNYKMDEIFRGRSTPLCDWLI
FRNVRSLLGGRTRLILCGGAPLSPSTQIFMNICFCCPVGQGYGLTETCGAGTISDVNDYSTGRVGAPLPC
SEIQLKNWEEGGYTVYDRPNPRGEIIIGGPNVSKGYFKNVTKTHEDYFVDKAGQWWFSTGDIGEFYPDGC
LKVIDRKKDLVKLQYGEYVALGKIESALKNSPIIENICAYANSDESYVICFVVPNQKHLMALAKEKNISG
TWSEICKNSDMEREVLNEIIEVAGQAKLETFEIPLKVHLSPEPWTPEMGLVTDAFKLKRKELKNYYEADI
ERMYGGK

>Cmictg24956 ACSl4

CCEIKLRDWEEGGYTCKDHPNPRGEILVGGPNVAMGYFKMNKSNDDFFLDSVGQRWFCTGDVGEFHPDGC
LQIIDRKKDLVKLQAGEYVSLGKVESALKSCSLIDNICAYANSEQSYVISFVVPNQKKLTILAEQKGIEG
VWEELCNNKVMEAEVLREIKEAAQSSKLEKFEVPMKVRLSPE

>Cmictg48116 ACSL6

GEYVAPEKIENIYIRSEPVAQVYVHGDSLQSFLVGIVVPDPEAADAWARKRGFEGPFVELCKAKGLKKAI
M

>Cmictg26023 ACSL4

CVAYGCRIGYSSPQTLSDQSTKIKKGSKGDCSVLKPTLMAAVPEIMDRIYKNVMSKVQEMSYVQRTLFKL
GYDYKLDQIKRGYDAPLCNLLLFKKVKALLGGNVRMMLSGGAPLSPQTQRFMNVCFCCPVGQGYGLTETC
GAGSITEVLDYSTGRVGAPLICCEIKL

>Cmictg18935
ACSL4ACFKYNFPLVTLYATLGEDSVAYGLNESEVTHLITSVDLIDTKLKNVLPKINKLKYLIYVDKKTI
STSGYPEKLRIHSMTCVEELGSKPENLNAHLNSPTPTDLAVVMYTSGSTGHPKGVMINHSNLIAGMTGQC
QRIPWLGPRDTYIGYLPLAHVLELTAEISCVAYGCR

>Cmictg34934 ACSL4

HIIYVDKKTINTSGYPDTLQIHNMAAVEEMGAKSETSEKASSRPAPSDLAVVMYTSGSTGRPKGVMMIHS
NLIAGMTGQCQRIPELGPKDTYIGYLPLAHVLELTAEISCVAYGCR>Cmictg79717 ACSL3
```

EIGLGIPGEPLFRSQGWFCTGDVGEFHHDGCLKVIDRKKDLVKLQAGEYVALGKVESALKNSPLIDNICV
YANSDQSYVICFVVPNQKQLLALAHQ

**Little skate**

>Lskctg17167ACSL1?

LANRLYDKIHSQANGFFKRKILQLASWRKIAELHKGIMRRDSIWDFLVFRKIQDSLGGQVKLMVIGAAPV
SDTVLDFVRAAMGCQFYEGFGQTECTAGCSMTIPGDWSAGHVGAPMPCNLVKLIDVHEMNYFSANGEGEV
CVKGPNVFSGYLKDPERTGEVLDETRWLYTGDIGKWLPNGTLKIIDRKKHIFKLAQGEYIAPEKIENIYV
RCELIAQVFVHGDGLQAFLVGIVVPDPEVVPEWANKWNLKGSYEELCKSQALKDAILANMLKVGKEAGLK
TFEQVKDIHLYPEMLTVESGLLTPTFKTKRLEMRKFFHKQIAELYAQNAV

>Lskctg11347ACSl1?

MQTVDLMTHLRMPELGEFQQYLRAVPTYSILGIGTIAALTMYWYTSKPKAQIPPCDLSMQSIEVDNFDHS
RKSVLLKEGQDLMTVYYQEVQTLYDVLKRGLIVSGNGPCVGCRKPNQPFQWISYREVSERAECIGSGFIR
RGYKGGNSEYIGIYAQNRPEWVIVEQACYAFSMVVVPLYDTLGIEAISFIINKADMEVVVCDTAGRARML
LNGVETGITAELKTIVVMESFGEDLVHRGIRHEVEVVSLAEIEKSGKEKKHRTSPPDPKDLAVICFTSGT
TGKPKGAMLTHKNIVANFSAFVKVTEGQWTASPSDIHISFLPLAHMFERLVQTVVLCHGGRIGFFQGDIR
LLMDDIKLLKPTIFPIVPR

>Lskctg31550ACSL5?

VEDMNYFAAKGEGEVCVKGTNVFKGYLKDPEKTLEAIDSDGWLHTGDVGKWLPNGTLKIIDRKKNIFKLA
QGEYIAPEKIENIYIQSGLVTQVFVAGDSLQS

>Lskctg34321ACSl5?

NSAACFRMMGSSISINVDDVSISYLPLAHMFERVVQTALYCSGSRVGFFQGNIQLLLDDMKTLQPTFFPV
VPRLLTRVYDKVQSS

>Lskctg43339ACSL5?

LKDDKVISHFYDDVKTLYEAFQRGLRVSGNGPCLGYRKPNQPYQWLSYKQVIDRAEHLGSGLLHRGCKPS
P

>Lskctg82046ACSL2?

ELAIVICDNMNKVKVLLGNCELKQTPSLSTIILMDPFDDALKERGTKCHVEVLTLKEVEDLGRENLQKPI
PPKPDDLSIVCFTSG

>LSKctg12475ACSL3?

MHFGSGLAVLGQRPKFNIAIFCETRAEWMIAAQSCFMYNFPLVTLYATLGGQAIAHGLNETEVTHIITSK
KLLQTKLKEILLHVPKLQHIIIVDDKTTAWADYPRGIMVHNMTGVEALGAKPENLNRARMRPTTLDIAVI
MYTSGSTGLPKGVMISHSNLVAGIAGMVARVPDLGPKDTYIGYLPLAHVLELSAETLCMACGCCIGYSSP
QTLADQSAKIKKGSKGDTTVLKPTLLAAVPEIMDRIYKNIVTKVEDMNSIQRTLFVVAYNYKQEQIAVGY
NTPICDWLIFRKIRLLLGGKTRLILCGGAPLSPSTQTFMNICFCCPVGQGYGLTETCGAGTISDVCDFTT
GRVGAPLPCCEITLKNWEEGGYTVYDKPHPRGEIVIGGPNVTLGYFKNKSKTMEDYFLDKNGQAWFCTGD
VGEFHHDGCLKVIDRKKDLVKLQAGEYVALGKVESALKNSPLIDNICVYANSDQSYVICF

>LSKctg14318ACSL4

MAKRIKAKATSDKPGSPFRSVDHFDSLASVDFPGLDTLDKLFDAAVKKFPEHDCLGTRERLSEENEVQAN
GKVFKKLILGEYRWLSYEDVGQRVNSFGRGLSALGLKPKDKIAIFCETRAEWMIAAQACFKFNFPLVTIY
ATLGEDAVTYGLNECEVTHLVTSMELLDTKLKSVLPKIRHLKHIIYVDKKTINTSGYPDTLQIHNMAAVE

```
EMGAKSETSEKASSRPAPSDLAVVMYTSGSTGRPKGVMMIHSNLIAGMTGQCQRIPELGPKDTYIGYLPL
AHVLELTAEISCVAYGCRIGYSSPQTLSDQSTKIKKGSKGDCTVLKPTLMAAVPEIMDRIYKNLMSKVQE
MNYLQRTLFKLGYDYKLEQIQRGYDAPLCNLVLFKKVKALLGGNVRMMLSGGAPLSSQTQRFMNVCFCCP
VGQGYGLTETCGAGSITEVSDYSTGRVGAPLICCEIKLRDWEEGGYTTHDQPNPRGEILIGGPNVAEGYF
KGNSTNDDFFEDSSGQRWFCTGDVGEFHSDGCLQIIDRKKDLVKLQAGEYVSLGKVESALKSCSLIDNIC
AYANSEQSYVISFVVPNQKKLTALAEQKRVQGTWEDICNDSKMEAEVLREIKEASASSKLEKFEVPIKVR
LSPEPWTPETGLVTDAFKLKRKELKNHYLNDIERMYGGK
```

**Coelacanth**

>LmaACSl1ENSLACG00000008655

```
MTTMMQAHDLLRHLRLPELGEVREYVRTLPTNTLMGFGAFAALTTYWYATRPKALKPPCDLSLQSVEIEG
SNGARRSALLETDQPLSFVYEDAQTMYELFQRGLRVSGNGPCLGFRKPNQPYEWISYQEVADGAEFLGSG
LIHRGCRPAPDQFIGIFAQNRPEWIITELACYTYSMVAVPLYDTLGAEAISYIINKAEIATVVCDEPDKA
RVLLENVEGGETPFLKMIVLMGHFENDLVERGKKCGVDILSMKAVEDLGKVNLCKPVLPKPSDLAVVCFT
SGTTGHPKGAMLTHGNIISNFSAFVKVTEKAVFPNTDDILISFLPLAHMFERIVEAVILCHGARIGFFQG
DIRLLMDDLKTLQPTIFPVVPRLLNRMFDRIFAQASTPIKRWLLEFAAKRKEAELRSGIVRSDSVWDKLI
FNKVQASLGGRVRLMVTGAAPISPSVLTFLRAALGCQFYEGYGQTEGTAGCTLTIPGDWTAGHVGAPMPC
NIIKLVDVEEMNYFSAKGEGEVCIKGTNVFQGYLKDPERTAEAFDKDGWLHTGDIGKWLPNGTLKIVDRK
KHIFKLAQGEYIAPEKIENTYVRSEPVAQVFVHGESLQAYLVGIVVPDPEVLPSWASKRGIEGSFAELCK
SRELKNTIIEDLVKLGKEAGLKSFEQVKDIMLYPEMFSIQNGLLTPTLKAKRPELRKYFKSQLDELYENI
KM
```

>LmaACSL2ENSLACG00000012732

```
IMESIEGLLDALHVPEYLGIPELEDASNFILSFSTPTLILIGAAVFALVYWLAMRPKATKLLCDLNNQSL
PVQGEPACRRSTLMKEDQYYCDDVKTAYEAFQRGVRISGDGPCLGFRKLGSPYNWISYKEVSSRAEFFGS
GLLHRGCSPSPDQFIAVFSQNRPEWVISKLACYTYSMVIAPLYETLGTEGLIHILNTTESSTVICDKPGR
AETLLSYVEGSKTPYVKTIILMEPFEDSLGGRGRSCGVDVLSMKDVEDLGKENWKAPMPPKPEDIAVICS
TSGTTGKPKLAMLSHKGMVCCISSVLSILQKKITPRAEDTILSFLPLAHVYELMAQLMFYCQGGRVGFYQ
GDIQLLMDDCRTLKPTFFPTVPRILNRMYDQIHSAMKSPVKRFLLHCAVWAKQVELRRGIIQNNSIWDWL
LFSRIQAILGGRVRIILTGSAPISSTVLSFFRTTLGCLIVEGYGQTECTGACTCSLPGDYTAGGHVGPPV
SWSTVKLEDVEEMNYFVSNGEGEICVKGPGVFLGYLKDTEKTAEAIDADGWLHTGDVGRWLPNGTLQIID
RKKHLFKLSQGEYIAPEKIENAYLRCTTVSQVFVHGESLQSFLVGIVVPDPEVLPSFAEKRGITGTYEEL
CRNPEVRKAVLEDMTRIGREAGLNSLEQVKTIYLHPEMFTIAKGMLTPTLKSIRAKLRNYFQQQINQLYT
NPSL
```

>LmaACSL6ENSLACG00000005834

```
FFFFFFLLHGTLFLMTEFAASALEKMQAQEILRSLRLPEFDEFSQFFRSMSAPTLVGIGAFAVVVAYWLA
SRPKAVRPPCDLAEQSQEVPGCDGAHRSVLGDTPQLLTHYYDDARTMYEVFQRGLHISENGPCLGFRRPK
QPYQWLSYKEVSNRAECLGSGLLQQGCKPSTDQFIGVFAQNRPEWIISELACYTYSMVVVPLYDTLGPGA
IRYIINTAEISTVICDKLEKARVLLEHVEKQETPGLKTVILMDPFHEDLVERGRKCGVHIQSMKEVEDLG
RANRRVPMPPRPEDLSIVCFTSGTTGNPKGAMLTHGNVVADFSGFLKVTEKVIFPRQDDVLISFLPLAHM
FERVIQSLWSPTCEDVHISYLPLAHMFERMVQSVVYCHGGRIGFFQGDIRLLSDDMKALRPTIFPVVPRL
LNRMYDKIFSQADTPFKRWLLEFAAKRKEAEVRNCIIRNDSLWDKLFFNKIQASLGGRVRMIVTGAAPAS
PTVLGFLRAALGCQVYEGYGQTECTAGCTFTTPGDWTSGHVGAPLPCNLIKLTDVQEMNYFATKGEGEIC
VKGPNVFKGYLKDLEKTAEALDENGWLHTGDIGKWLSNGTLKIIDRKKHIFKLAQGEYIAPEKIENIYIR
SEPVAQLYVHGDSLQSCLVGIVVPDPEVMPTWAKKRGFEGSYAEMCKNKELKIAIMQDMVRLGKESGLHS
FEQVKDIYIHSEMFSVQNGLLTPTLKAKRTELREYFKKQIEELYANVSM
```

```
>LmaACSL5ENSLACG00000010845

MNFLLHFLFSPLSTPALIGIFSFGAAIFIWLITRPKPVRPPVDLNNQSVGLKEGARRSALLKSDKLISYY
YDDARTLYEIFQRGLHISGAGPCLGYRKPKQPYQWLTYKQVSDRAEYLGSGLLHRGCRPAPDQFIGIFAQ
NRPEWIISEFACYTYSMVAVPLYDTLGPEALIFIINRAEISTVICDKPSKAITLLDNCEERQTPGLNTII
LMDPFEDDLKEKGAKFGVEILTLQQVEDLGKENLRKPIPPKPEDLSIVCFTSGTTGDPKGAMLTHENVVA
DSAAFIKSLESTFAPLQEDISISYLPLAHMFERVVQTIMYSSGAKVGFFQGDIRQLPDDMMALQPTVFPV
VPRLLNRIYDKVQSGAQTPFKKWLLNFAVARKHAEVKQGIIRNDSIWDNLIFHKVQATIGGRARVMVTGA
APISPTVLKFLRAVLGCQIFEAYGQTECTAGCTFSIAGDWTTGHVGAPLPCNLVKLVDVEEMNYFAVNGE
GEVCIKGTNVFKGYLKDSEKTAEALDADGWLHTGDVGKWMSDGTLKIIDRKKNIFKLAQGEYIAPEKIEN
VYIRSSPVAQVFVHGDSLQSCLVGIVVPDPEVLPDFAAKLGQNGSYEELCKNPVVRKAILEDMVKLGQKA
GLKSFEQVKEIYIYTEMFTIENGLLTPTLKAKRAELSKYFKGQIDSLYASMQG
```

```
>LmaACSL3ENSLACG00000003984

MKLEEDLHPVCLYIIHFLIKVYTCITFLPWYYFSGASQNLAKAQQVKARPVQNCPGKPYRSVNSLHCLAS
VLYPGCDTLDKVFEYAKNKFKDEDCLGTREVLNEDDEIQPNGRVFKKVILGRYNWLSYEDAYVKAALFGN
GLAQLGQMPRCTIAIFCETRAEWMIAAQACFMYNFPLVTLYATLGSRAIAHGLKESAVSHIITSKELLQT
KLKPIVTDVPQLKHIIIVDEKPTAWTGYPSGITVHSMAAVQALGTKPEDLNVPHRRPKPSDIAVIMYTSG
STGLPKGVVISHSNLIAGITGMAERIPNLGVKDTYVGYLPLAHVLELSAELVCLSHGCRIGYSSPQTLAD
QSTKIKKGSKGDTSVLKPTLMAAVPEIMDRIYKNVMNKVNEMNRFQRNLFILAYNYKMDQISKGCSTPLC
DSLVFRKVRSLLGGKTRVILSGGAPLSPATQRFINICFCCPVGQGYGLTETCGAGTISEVWDYSTGRVGA
PLVCCEIKLKDWEEGGYYNTDKPHPRGEILIGGQNVTMGYFMNAEKTKEDFLVDLDGQRWFCTGDIGEFH
PDGCLKIIDRKKDLVKLQAGEYVSLGKVEAALKNCPLIDNICAYANSDQSYVIGFVVPNQKELLYLSAQK
GIKGTWEEICNNHEMEKEVLKVIADVAISAKLEKFEVPVKVRLSAEPWTPETGLVTDAFKLKRKELKSHY
QEDIERMYGGK
```

```
>LmaACSL4ENSLACG00000002977

TMKLKVLKVSSILLLPVYVLMFVYTILTFIPWYFLTNAKKKKAMAKRIKAKPISEEPGSPYRSVDHFNSL
ATIGIPGADTLDKLLDQAVVKFGKKDCLGTREFLSEENEVQPNGKVFKKLILGEYKWISYEEVHKQVTHF
GSGLAALGQKPKNTIAIFCETRAEWMISAQACFKHNFPVVTLYATLGEQAVAYGLKESQVTHLITSIELL
ETKLKRVLSNIPNLKHVIYVDKKNFDKSGYPAGIQIHSMASVTELGAKPENLNSPITHPVLSDLAVVMYT
SGSTGLPKGVMMVHSNLIAGMAGQCNKIPELGPKDTYIGYLPLAHVLEMTAEISCIAYGCRIGYSSPQTL
SDQSTKIKKGSKGDCTVLKPTLMAAVPEIMDRIYKNVMSKVQEMSYVQKTLFKLGYDYKLEQIKRGYDAP
LCNLFLFKKVKALLGGNVRMMLSGGAPLSPQTQRFMNICFCCPVGQGYGLTETCGAGTITEATDYSTGRV
GAPLICCEIKLRDWPEGGYTTRDKPNPRGEIVIGGPNVSMGYFKSEEKTSEDFIIDRNGQRWFCSGDIGE
FHPDGCLQIIDRKKDLVKLQAGEYVSLGKVEAALKNCALIDNICVYANSYQSYVISFVVPNQKKLMALAQ
HKGIEGSWEELCNNPTMEAEVLKAITDVASSVKLERFEIPVKVRLSPDPWTPETGLVTDAFKLKRKELKN
HYLNDIERMYGGK
```

**Spotted Gar**

```
>Loc_ACSL2_ENSDARP00000053749_1

SLSPSSLLGLGALASLTAYWLATRPRPIRPPCDLHAQSVPVQGDPSCRRSALLQDEMLLEFYYEDTRTAY
EMFQRGLRVSGDGPCLGFRKPSEPYKWISYREVCEQAQALGSGLLARGCQPNPQQFIGIFAQNRPEWVIA
ELACYSFSMAVVPLYDTLGQEAMVHILNIAEITMVICDKPEKAESLLSHKEQTLAPLLGSIVLMTPCSAA
LLERAKKCGIEILQFSELMVSQRECVLVILPPKPEDLAVVCFTSGTTGKPKGAMITHGNIASNTSSVIKI
LEGSFVIRQEDISISYLPLAHMFERMIQVSMFCHGARVGFYQGDLSLLMDDIKTLRPTFFPVVPRLLNRI
```

```
YDKILASVSSPLKRALLHYAVRRKQAELSSGVVRNNSVWDRLIFNKIQASLGGNLRFILTASAPISPTVL
SFLRATLGCLIFEGYGQTECTAGCTFSMPGDWTAGHVGAPLPCAMVKVTDIPEMSYYSHNGEGEICIKGH
SVFRGYLRDHERTAEALDPEGWLHTGDVGKWLPNGALQIIDRKKHIFKLSQGEYIAPEKIENVYIRSAPV
LQVFVHGDSLQSHLVGIVVPDPEVFTDWAKERGIVGSYEELCRNPDVKKAVLEDMTVIGKEAGLKSFEQV

>Loc_ACSL6_ENSDARP00000039916_1?

PPQPDDLSIVCFTSGTTGNPKGVMLTHGNVVADFSGFLKVTEKVIFPRQDDVLISFLPLAHMFERLIQSV
VYCHGGRIGFFQGDIRLLSDDMKALCPTIFPVVPRLLNRMYDKIFSQASNPVKRWLLEFAARRKSAEVHS
GIIRNDSVWDKIFFSKIQASLGGRVRMIVTGAAPASPTVLGFLRAALGCQVYEGYGQTECTAGCTFTTPG
DWTSGHVGAPLPCNLIKLVDVADMNYFASKGEGEVCVKGPNVFKGYLKDQEKTAEALDEDGWLHTGDIGK
WLPNGTLKIIDRKKHIFKLAQGEYIAPEKIENIYIRSEPVAQLYVHGDSLQSCLVGIVVLDPEVLLDWAR
KRGIEECFGDLCKNKEVKKAVMDDMVRLGKASGLHSFEQVKDIYIHNELFSIQNGLLTPTLKAKRSELRE
FFKEKIEHMYANISM

>Loc_ACSL5_ENSDARP00000123999_1

MSCILQFLFSPLPTPAIAGLFTFGAGILIWLVTRPKPIKLPVDLNRQTVGIKDGARRSALLKDDKLMSYY
YEDARTLYEVFQRGLHVSGNGPCLGYRKPRQPYQWLKYKQVSDRAEYLGSGLLHRGLKPSPDQFIGIFAQ
NRPEWIISELACYTYSMVAVPLYDTLGPEALVYIVNKAEISTVICDKPDKAAILLTNCEKGLTPVLNTIV
LMDPFSTDLKDRGMNCGVEILALKEVEV*GSKFCFATEIPPKPEDLSIVCFTSGTTGDPKGAMLTHENVV
ADAAGFIKSTESAFGPVPQDVSISFLPLAHMFERVVQTVMYSSGAKVGFFQGDIKLLPDDMKALRPTVFP
VVPRLLNRVYDKVQSGAQTPFKKWLLNFAIERKHSEVKQGIIRNDSIWDKLIFHKVQETMGGRVRVMVTG
AAPISPSVLSFLRACLGCQIFEAYGQTECTAGCTFSMPGDWTTGHVGVPVPCNIVKLVDVEEMNYFAANG
EGEVCIKGRNVFKGYLKDPEKTSEAIDENGWLHTGDIGKWLPSGVLKIIDRKKNIFKLAQGEYIAPEKIE
NVYYVRSEPVAQVFVHGDSLQSCLVGIVVPDIEVLPDFAVKLGIKGSYKELCINKEIKKAILADMVRLGKQ
AGLKSFEQVKDLYLYPEQFTIENGLLTPTLKAKRAELTKFFKDQIDSLYA

>LocACSL4_ENSDARP00000010373_1

EVHSILLLPVHLFMWLYTLLTFIPWYFLTDTRKKKTMAKRIKAKSTTGKAEGPYRSVDHFQSLATMDFEG
KDTLDKLFDHAVQRFGKADCLGTREVLTEENETQPTGKVFKKLILGEYKWLTYEDVNRQVTLFGSGIAAL
GQQPKNTIAIFCETRAEWMIAAQACFRRNFPVVTLYATLGEDAVAYGLNESGVTHLITSVELLETKLKKV
LSEIKNLKHIICVDKKNASKTGYPEGLHIHSMESVQELGGKPENLSVTPCHPQPTDLAVVMYTSGSTGRP
KGVMIIHSNLIAGMTGQCERIPGLGPKDTYIGYLPLAHVLEMTAEISCVTYGCRIGYSSPQTLSDQSTKI
KKGSKGDCSVLKPTLMAAVPEIMDRIYKNVMSKVQEMSYVQRTFFKLGYNYKLEQIKMGYDAPLCNLLFK
KVKALLGGSVRMMLCGGAPLSSATQRFMNICFCCPVGQGYGLTETCGAGTITEVADYSTGRVGAPLICCE
IKLRDWAEGGYTNQDVPHPRGEILIGGPNVTMGYYKNGQINEDFFVDENGQRWFCTGDIGEIHADGCLQI
VDRKKDLVKLQAGEYVSLGKVESALKNCSLIDNICAYANSDQNYVISFVVPNQKRLTALANKQGISGAWE
DICNHPAMESEVLKEIKEVATSIKLQRFEIPVKVRLSPEPWTPETGLVTDAFKLKRKELKNHYLNDIERM
YGGK

>LocACSL3ENSDARP00000042883_1

LLSGASQNLERAKRVKARPVNNQPGGPYRSVNSMSCLASSLYPGCDTLDKVFEYAKNKFSTNHCLGTREL
LSEEDEVQPNGKVFKKVILGEYRWLSYKETHLAAARFGSGLAALGQKPKSTIAIFCETRAEWLITAQACF
MYNFPLVTLYATLGGPAIVHGLNETEVTHIITSKDLLQSRLKAILLEVPRLQHIIIVDDKSSTWIDYPRG
ITIHNMAEVQALGSKEENMSKPRCQPAPSDIAVIMYTSGSTGIPKGVMISHSNLIAGITGMAERIPDLGE
NDTYIGYLPLAHVLELSAELVCVSHGCRIGYSSPQTLADQSTKIKKGSKGDTSVLKPTLMAAVPEIMDRI
YKNVMRKVEEMNSVQRTLFVLAYNYKMEQISKGYSTPLCDSFVFRKVRSLLGGRTRVLLSGGAPLSAATQ
RFMNICFCCPVGQGYGLTETCGAGTISEMWDYSTGRVGAPLVCSEIKLKSWEEGGYYCTDKPNPRGEILV
GGPNVTMGYYKNEAKNKDDFFVDEKGQRWFCTGDIGEFHPDGCLKIIDRKKDLVKLQAGEYVSLGKVEAV
LKNCPLIDNICAYANSDQSYVIGFVVPNQKQLMALAEHKKVGGTWEEICNNSEMEKEVLRVMAEAASGAK
LEKFEIPMKIRLSAEPWTPETGLVTDAFKLKRKELKTHYQDDIERMYGGK
```

**Small spotted catshark**

```
>SSCctg18625ACSL1?

VVFCDTAVKAEAVLLGVEKGQTPDIKTIIIMDPFGAELKNRGKAYGVEIVSLNIIETAGREMKRDPRLPQ
PSDLAVICFTSGTTGNPKGAMLTHRNIVSNFSAFVKVTEGQWVASPSDIHISFLPLAHMFERLVQIVVLC
HGARIGFFQGDIRLLMDDIKVLQPTIFPVVPRLLNRMYDKVHSQANSFLKRQILALATWRKTAELRKGIM
RRNSIWDKLVFHKIQESLGGKVRLMVTGAAPVSDTVLTFIRAAVGCQFYEGYGQTECTAGCSMTIPGDWS
AGHVGAPMPCNLIKLVDIEEMNYFAREGEGEICIKGSNVFLGYLKDPEKSAEALDQRGWLHTGRGQHGNG

>SSCctg67516ACSL3

MKLKKGVNPIFSLFLQCVIVVCNLLFILPLQLFAGSRRRPSIRAKSISNHPAGPYRCVESLDRLLASLYP
GADTLDKIFQFATDGFRHKNCLGTREILSEEDEIQPSGRVFKKLILGHYKWLTYDEVYRRVAHFGSGLAM
LGQRPKANIAIFCETRAEWMIAAQSCFMYNFPLVTLYATLGGQAIAHGLNETEVTHIITSKKLLQTKLKE
ILLNVPKLQHIIIVDDKPTAWSEYPRGIMVHNMAAVEALGSKPENLNRVRAKPTSLDIAVIMYTSGSTGL
PKGVMISHSNLIAGIAGMCSRVPDLGPKDTYIGYLPLAHVLELSAETLCMACGCGIGYSSPQTLADQSAK
IKKGSKGDTTVLKPTLLAAVPEIMDRIYKNIVTKIEDMSSMQRTLFVVAYNYKMEQMSLGCSTPICDWLI
FGKIRSLLGGKTRLILCGGAPLSPSTQTFMNICFCCPVGQGYGLTETCGAGTISDVFDYTTGRVGAPLPC
SEITLKNWEEGGYTVYDKPHARGEILIGGPNVTLGYFKNKSKTLEDYFVDKDGQAWFCTGDIGEFQDDGC
LKVIDRKKDLVKLQAGEYVALGKVESALKNSPLIDNICAYANSDQSYVICFVVPNQKQLLALAQQKGIIG
SWNDICNRPEMEKEVLREITDAAAVARLEKFEIPLKVRLSPEPWTPETGLVTDAFKLKRKELKTHYLADI
ERMYGGK


>SSCctg66619ACSL4

MAKRLKAKATSEKPGSPFRSVDHFDSLAKMDFPGLDTVDKLFEAAEKKFRKQHCLGTRELLSEENEVQVN
GKVFKKLILGEYKWLSYEEVNQHVNCFGSGLTALGLKAKDMIGIFCETRAEWMIAAQACFKYNFPLVTIY
STLGEDAVAYGLNESEITHLITSAELLDTKLKKVLPKIQMLKHIIYVDNKVINTSGYSETLQIHSMESVE
ELGAKPENMDITPSRPVPSDLAVVMYTSGSTGHPKGVMMIHSNLIAGMTGQCQRIPGLGPKDTYIGYLPL
AHVLELTAEISCVAYGCRIGYSSPQTLSDQSTKIKKGSKGDCSILRPTLMAAVPEIMDRIYKNVMSKVQE
MNYVQRTLFKLGYDYKLEQIKRGYDAPLCNMLLFKKVKSLLGGNVRMMLSGGAPLSPQTQRFMNVCFCCP
VGQGYGLTETCGAGSITEVLDYSTGRVGAPLICCEIKLRDWEEGGYTTNDQPHPRGEILIGGPNVAMGYF
KLNKSSHDFFEDNTGQRWFCTGDVGEFHPDGCLQIIDRKKDLVKLQAGEYVSLGKVESALKSCSLIDNIC
AYANSEQSYVISFVVPNQKKLTALAEQKQVQGTWEEICNDSKMEAEVLREIKEASASSKLEKFEVPVKVR
LSPEPWTPETGLVTDAFKLKRKELKNHYLNDIERMYGGK

>SSCctg94905ACSL1?

MQASELLTQLRIPEFGEVRRFICSLPPSTLIGIGTIAALVAYWYATRAKAQKPPCDLSKQSVEVEGGERA
RRSVLLKSDEPMVFYYLDVKTLYDVLKRGLWVSDNGPCLGFRNPDQPYQWLSYREVITRAEFVGSGLFTR
GYKPGNDQFIGIFAQNRPEWVIIEQACYTYSMVVVPLYDTLGDEAISYILNKADIAVVFCDTA

>SSCCtg81762ACSL6?

MDPFEMDLVEMGNDCGVQILALQEVENLGRVNRQTPVPPRPEDLSIVCFTSGTKGKPK


>SSCctg63419 possible Acsl2
```

```
CFRSLDCSKKNCRSSPDQFIGVFAQNRPEWIIAELACYTYSMVIVPLYDTLGPEAIRYIINTAEISTVIC
```

**Lamprey**

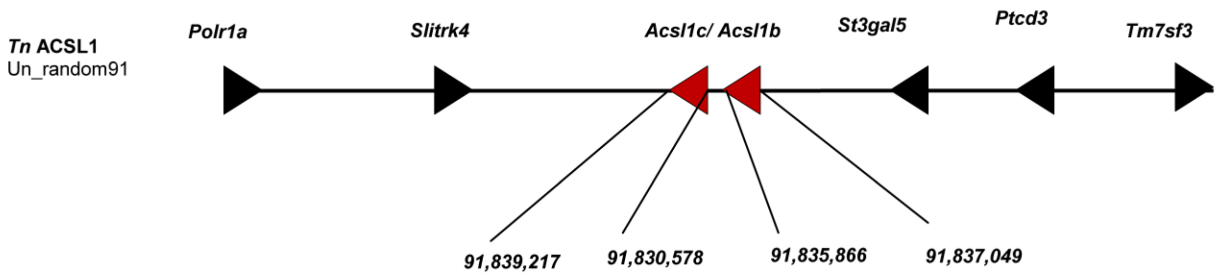
```
>PmaASCL3ENSPMAG00000005133?

QAILMQVPRLRYVVLVDGTASGVAGVKLPRGIEVMGMSDVEELGDKPQHRQRARDRPGPRDLAVVMYTSG
STGIPKGVRIAHSNLVAAITGMYLRINDICGDDTYIGYLPLAHVLELGAEMVCLSRGCRIGYSSAQTLTD
QSTRIKKGSQGDVTILRPTLMAAVPEIMERIYKGVMGKVQCMSLLQRIIFKLAYNYKLEQLERGFDTPLC
NRLVFNKLCALLGGNVRLLLSGAAPLSPRTQRFMNVCFCCPVGQGYGLTETCAAGTIAELQDYSTGHVGA
PLSCCEIQLRNWEEG

>PmaACSL1ENSPMAG00000008135

MQSAQEVLKTLRVPELDEVRQYVRSLPAPALMGLGALGTMTAYWLATRPRALSAPCDLSKQSLPVKGREY
QRRSPLVPDDDTFFTYFYEDARTGYEVFQRGLRISNNGPCLGYRKPNHPYEWISYKETSDRAEYLGSGFL
HLGAKPSSEQVIGIFSQNRPEWIIAEQACYTYSLVVVPLYDTLGRESIDYIINQAEISMVVCDKLEKVKG
LLESIEEGAIRIVKTIVVMDPFDGVMEQRARKCGIDLILFRELEVIGKTNHREPIPPQPDDIAIICYTSG
TTGNPKGAMLTHKNMISDFSAFLAITKDTFLPNTDDVLISFLPLAHMFERLVEASILCNGGRVGFYQGDI
RLLMDDMKVLQPTVFPVVPRLLNRIHDKVLSGAKSHFKRWLLEFAVSRKIAELRCGVVRKDSIWDKLIFH
RVQASMGGRVRFMVTGAAPISASILTFLRAILGCQVYEGFGQTECTAGCTFTMPSDSTAGHVGPPMPCNH
IKVVDVAEMNYFAANGEGEVCVYGTNVFKGYLKDPTRTAEALDEDGWLHTGDIGKWLPNGTLKIVDRKKH
IFKLSQGEYIAPEKIETVYVRSEPVAQVFVHGDSLQSCLVCIVVPDMEVLPSWVQKRGIKCNPNSVFINK
DVRAAILHDMVRLGKEAGLKSFEQVRAVHLHSDLFSIENGLLTPTFKVKRAEVCKFFHSEIDSLYAGITV


>PmaACSL1ENSPMAG00000004625?

SLQDCGRMHKRTTLPPKPEDLAIVCFTSGTTGNPKGAMLTHRNIVSDMSGFLKVTESLFLPETSDIAISY
LPLAHMFERLVQATLFCHGASIGFFQGDVRLLLDDMQALRPTVFPVVPRILNRMYDKVLGSTRTPFRRKL
LEFGARRKMAELQRGVVRRNSLWDWLVFRPMQLSVGGRVRMIMTGAAPISPGVLNFLRVVMGCQMYEGYG
QTECTAGCTLTLPGDWKAGHVGAPMPCNYIKLHDVKDMEYYTSQGKGEVCVKGPNVFKGYLKDAEKTAEA
VDRDGWLHTGDIGQWMPNGTLKIIDRKKHIFKLAQGEYIAPEKIENVYSRCEPVAQVYVHGDSLQACLVA
VVVPDPEILCCWIRKKGIVGTYSELCRNKEVRQAILEDMQRLGKESDLQPFEQVKDVHLHNEMFSIENGL
LTPTFKAKRTELRTHFRSVISSMYQNVKA


>PmaENSPMAG00000005099ACSL1?

VEEYIPWRTQMLSIGGLLENVWGSMPPSAWLGVGTASLFTGYWYLSRPRPICPPCQLTQQSVEVEGQDGV
RRSALLKKDKLLEFYYEDAKTMYEVFHRGMRVSKNGPCLGYRKPKKPYQWMSYKEVAERSEWFGSGLIHK
GCRPATDQFIGVFAQNRPEWIITEQACYMYSMVVVPLYDTLGQEAIRFIINRAAEIAVVVCDTVERARVL
LRGVENRETPGLRTVVVMETPDLGLIQWGITCRVDVLSFQYIE
```

Additional file 2

## Partial *Acsl* gene annotations in green spotted pufferfish

### A- Genomic location of the partial sequences of *Acsl*1 from green spotted pufferfish

*Acsl1a* - ENSTNIG00000018054 (complete sequence)
*Acsl1b*- ENSTNIG00000000345 (partial sequence)
Acsl1c- ENSTNIG00000010115 (partial sequence)



### B- Alignment of the partial *Acsl1b* and *Acsl1c* sequences and complete *Acsl1a* sequence

Highlighting in grey marks the overlapping region between the two partial sequences

```
ACSL1a_ENSTNIG00000018054      MQTPEALKQFWIPELDNIQHFLGGMSGNALVGMGVLVALTTYWLASRHRA
ACSL1b_ENSTNIG00000000345      IFIMDLVYRLGLLSLDSVTQYVRSVSTPVWVGTGLVAAATTYLLTARPKA
                               :   : : :: : .**.: ::: .:*  . ** *::.* *** *::* :*

ACSL1a_ENSTNIG00000018054      VKQRVDFSRQSVELPGGEGIRRSVLVENDQLITHYYDDARTFYELFLRGL
ACSL1b_ENSTNIG00000000345      LPPICDLDMQSIEIPGGELARRSALQNGDAYTKCYYDDARTMYESFLRGL
                               :    *:. **:*:****  ***.* :.*   . *******:** *****

ACSL1a_ENSTNIG00000018054      RESNNGPCLGSRKLNHPYEWQSYQEVVADRAKHIGSALLNKGHSHTGDKF
ACSL1b_ENSTNIG00000000345      RVSNDGPCLGSRKPKQPYEWLSYSEVK-ERAENLGSAFLHRGHSKTKDPH
                               * **:******** ::**** **.**  :**:::***:*::***:* * .

ACSL1a_ENSTNIG00000018054      IGIFSLNRPEWTISELACYTYSLVAVPLYDTLGREAIGYIIDKATISTLI
ACSL1b_ENSTNIG00000000345      IGIFSQNRAEWTISELACYTYSLVSVPLYDTLGTEAIIYIVEKASISTIV
                               ***** **.***************:******** *** **::**:***::
```

```
ACSL1a_ENSTNIG00000018054     CDLPEKAWMVLDCINGKGKSVKRIVIMGPFQSELVERAEECDIEIISFED
ACSL1b_ENSTNIG00000000345     CDLSSKVDLLLSCLEDKKHAVKTVVLMEKPSVELVSRAKRSGIDVISVEE
                              ***..*. ::*.*::.* ::** :*:*   . ***.**:...*::**.*:


ACSL1a_ENSTNIG00000018054     FEALGQDTVMEPVPPAPEDLALVCFTSGTTGKPKGAMLTHGNIIANTAAF
ACSL1b_ENSTNIG00000000345     MEALGKANRQPPVPPKPEDMAVICFTSGTTG-------------------
                              :****: .   **** ***:*::********

ACSL1a_ENSTNIG00000018054     LKLTEKDCMLCVHDIHISYLPLAHMLERVIHGVVLVHGGRVGFFQGDIRL
ACSL1b_ENSTNIG00000000345     --------------------------------------------------

ACSL1a_ENSTNIG00000018054     LMDDLQTLKPTVFPMVPRLLNRMCDKIFSQADTPLKKWLLRLAFSRKIAE
ACSL1b_ENSTNIG00000000345     --------------------------------------------------

ACSL1a_ENSTNIG00000018054     LNQGVVRQDTIWDRLIFKKVQANTGGRVRMMITGAPPVCPKNLTYINITT
ACSL1b_ENSTNIG00000000345     --------------------------------------------------

ACSL1a_ENSTNIG00000018054     MLQLYEGYGQTECTAGCSMSLPGDWIAGAVGPPVPCNDIKLVDVAEMNYF
ACSL1b_ENSTNIG00000000345     --------------------------------------------------

ACSL1a_ENSTNIG00000018054     AANGEGEVCAKGTNVFKGYLGDAEKTAEALDEDGWLHTGDIGKWLPNGTL
ACSL1b_ENSTNIG00000000345     --------------------------------------------------

ACSL1a_ENSTNIG00000018054     KITDRKKNIFKMAQGEYIAPERIEMIYNRSEPVAQIFVHGDSLKACLVAI
ACSL1b_ENSTNIG00000000345     --------------------------------------------------

ACSL1a_ENSTNIG00000018054     VVPDSETLPDWIKKKGIEGPPTGLCKNQDVKRAIQEDILRLGREAGLKSF
ACSL1b_ENSTNIG00000000345     --------------------------------------------------

ACSL1a_ENSTNIG00000018054     EQVKDITLHPEMFSIQNGLLTPTLKSKRVELRRYFRKQIDEMYAKIKR
ACSL1b_ENSTNIG00000000345     ------------------------------------------------
```

---

```
ACSL1c_ENSTNIG00000010115     --------------------------------------------------
ACSL1a_ENSTNIG00000018054     MQTPEALKQFWIPELDNIQHFLGGMSGNALVGMGVLVALTTYWLASRHRA

ACSL1c_ENSTNIG00000010115     --------------------------------------------------
ACSL1a_ENSTNIG00000018054     VKQRVDFSRQSVELPGGEGIRRSVLVENDQLITHYYDDARTFYELFLRGL

ACSL1c_ENSTNIG00000010115     --------------------------------------------------
ACSL1a_ENSTNIG00000018054     RESNNGPCLGSRKLNHPYEWQSYQEVVADRAKHIGSALLNKGHSHTGDKF

ACSL1c_ENSTNIG00000010115     --------------------------------------------------
ACSL1a_ENSTNIG00000018054     IGIFSLNRPEWTISELACYTYSLVAVPLYDTLGREAIGYIIDKATISTLI

ACSL1c_ENSTNIG00000010115     --------------------------------------------------
```

```
ACSL1a_ENSTNIG00000018054       CDLPEKAWMVLDCINGKGKSVKRIVIMGPFQSELVERAEECDIEIISFED

ACSL1c_ENSTNIG00000010115       -----------------------CFGTVVSGDPKGAMLTHENIVSNCSAV
ACSL1a_ENSTNIG00000018054       FEALGQDTVMEPVPPAPEDLALVCFTSGTTGKPKGAMLTHGNIIANTAAF
                                                       ** : .:*.******** **::* :*.

ACSL1c_ENSTNIG00000010115       IKVTEVSCPFCSSDTHMSYLPLAHMFERIVQGVVLVHGARIGFFQGDIRS
ACSL1a_ENSTNIG00000018054       LKLTEKDCMLCVHDIHISYLPLAHMLERVIHGVVLVHGGRVGFFQGDIRL
                                :*:** .* :*  * *:********:**:::*******.*:******** 

ACSL1c_ENSTNIG00000010115       LSDDLCALKPTVFPVVPRLLNRMYDRIFGQANSTVKRWLLGFAFRRKEAE
ACSL1a_ENSTNIG00000018054       LMDDLQTLKPTVFPMVPRLLNRMCDKIFSQADTPLKKWLLRLAFSRKIAE
                                * *** :*******:******** *:**.**::.:*:*** :** ** **

ACSL1c_ENSTNIG00000010115       LRRGIMRRDSIWDRLIFRKVQASLGGRVRFMITGAAPISPAVLTFLRVAM
ACSL1a_ENSTNIG00000018054       LNQGVVRQDTIWDRLIFKKVQANTGGRVRMMITGAPPVCPKNLTYINITT
                                *.:*::*:*:*******:****. *****:*****.*:.*  **::.:: 

ACSL1c_ENSTNIG00000010115       GCQFFEGYGQTECTAGCTMTLAGDWTAGHVGPPLPCNSVKLVDVAEMNYL
ACSL1a_ENSTNIG00000018054       MLQLYEGYGQTECTAGCSMSLPGDWIAGAVGPPVPCNDIKLVDVAEMNYF
                                  *::************:*:*.*** ** ****:***.:**********:

ACSL1c_ENSTNIG00000010115       AANGEGEVCVKGPNVFQGYLHDPEKTAEAIDAHGWLHTGDIGKWLPNGTL
ACSL1a_ENSTNIG00000018054       AANGEGEVCAKGTNVFKGYLGDAEKTAEALDEDGWLHTGDIGKWLPNGTL
                                *********.**.***:*** *.******:* .*****************

ACSL1c_ENSTNIG00000010115       KIIDRKKHIFKLAQGEYIAPEKIENVYVRSSAVAQVYVHGDSLQAFLVAV
ACSL1a_ENSTNIG00000018054       KITDRKKNIFKMAQGEYIAPERIEMIYNRSEPVAQIFVHGDSLKACLVAI
                                ** ****:***:*********:** :* **..***::******:* ***:

ACSL1c_ENSTNIG00000010115       VVPDPDFLCGWAKKTLGLRGSYEDLCSKESVWVSVFERCVIVIEDACVVL
ACSL1a_ENSTNIG00000018054       VVPDSETLPDWIKK-KGIEGPPTGLCKNQDVKRAIQEDILRLGREAGLKS
                                ****.: * .* **  *:.*.  .**.::.*  :: *  : : .:* :  

ACSL1c_ENSTNIG00000010115       CVQVKAIAIHPELFSVENGLLTPTLKAKRNEMRQFFRPQLDHLYASIKM
ACSL1a_ENSTNIG00000018054       FEQVKDITLHPEMFSIQNGLLTPTLKSKRVELRRYFRKQIDEMYAKIKR
                                  *** *::***:**::*********:** *:*::** *:*.:**.** 
```

Additional file 3

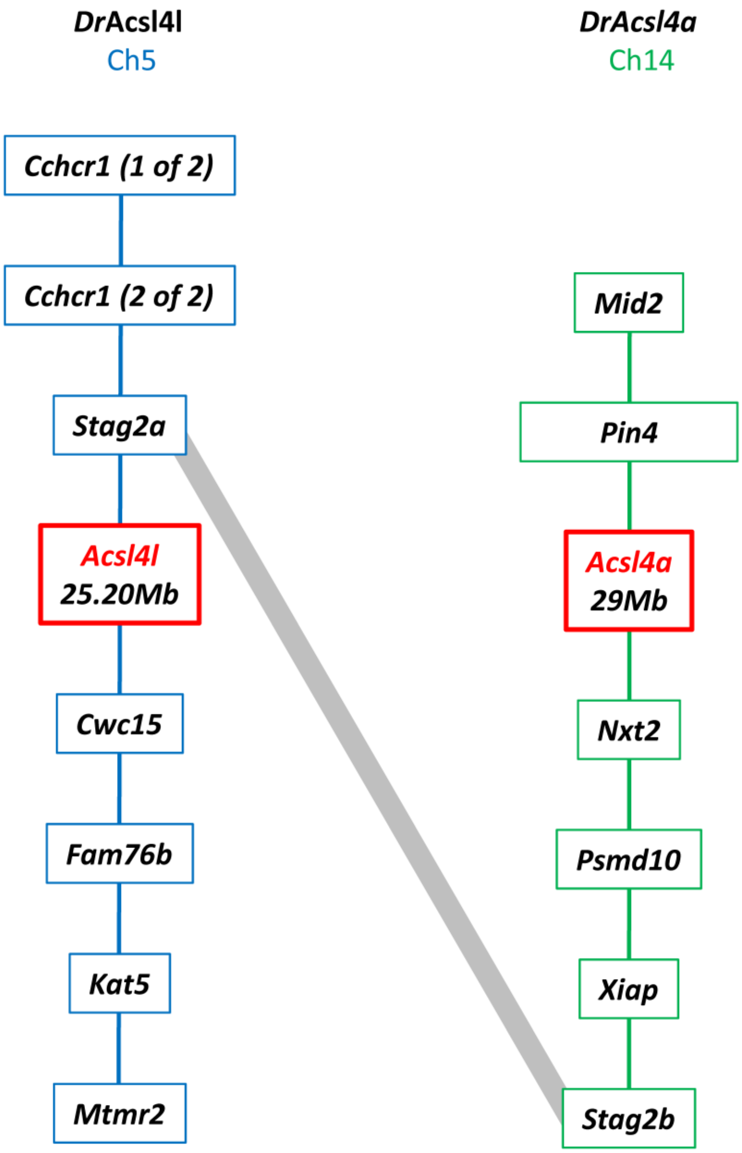
## NCBI accession numbers and ensemble gene ID numbers of the final set of ACSL sequences gathered after extensive database searches.

| | ACSL1 | ACSL3 | ACSL4 | ACSL5 | ACSL6 | ACSL2 |
|---|---|---|---|---|---|---|
| ***Hs*** | NP_001986.2 | NP_004448.2 | NP_075266.1 | NP_057318.2 | NP_056071.2 | na |
| ***Mm*** | NP_032007.2 | NP_083093.2 | NP_997508.1 | NP_082252.1 | NP_659072.3 | na |
| ***Md*** | XP_001363547.1 | XP_001365624.2 | XP_001363499.1 | XP_003339985.1 | XM_001371426.1 | na |
| ***Xt*** | NP_001006830.1 | NP_001090679.1 | XP_002938836.1 | NP_001011069.1 | NP_001072330.1 | na |
| ***Ac*** | XP_003221679.1 | XP_003218347.1 | XP_003223418.1 | XP_003223478.1 | XP_003217452.1 | na |
| ***Gg*** | NP_001012596.1 | XP_422625.2 | XP_420317.2 | NP_001026408.1 | ENSGALG00000006644 | na |
| ***Oa*** | XP_003429626.1 | ENSOANG00000013598 | XP_001507836.2 | XP_001513244.1 | ENSOANG00000009215 | na |
| ***Tr*** | ENSTRUG00000017576 | ENSTRUG00000009826 | ENSTRUP00000006243 | ENSTRUG00000007791 | ENSTRUG00000004657 | ENSTRUG00000004367 |
| | | ENSTRUG00000013537 | | | | |
| ***Ol*** | ENSORLG00000018806 | ENSORLG00000009215 | ENSORLG00000008040 | ENSORLG00000011808 | ENSORLG00000001111 | ENSORLG00000011037 |
| | ENSORLG00000008655 | ENSORLG00000015909 | | | | |
| ***Ga*** | ENSGACG00000017662 | ENSGACG00000000252 | ENSGACG00000018503 | ENSGACG00000003579 | ENSGACG00000020877 | ENSGACG00000010837 |
| | ENSGACG00000018774 | ENSGACG00000014028 | | | | |
| ***Tn*** | ENSTNIG00000018054 | ENSTNIG00000015280 | ENSTNIG00000015788 | ENSTNIG00000014415 | ENSTNIG00000016086 | ENSTNIG00000010817 |
| | | ENSTNIG00000014309 | | | | |
| ***Dr*** | NP_001003569.1 | ENSDARG00000032079 | ENSDARG00000004078 | ENSDARG00000075931 | XP_001920939.3 | ENSDARG00000078399 |
| | CAX14650.1 | ENSDARG00000014674 | AAH91952.1 | | | |

**Hs-** *Homo sapiens*; **Mm-** *Mus musculus*; **Md-** *Monodelphis domestica*; **Xt-** *Xenopus tropicalis*; **Ac-** *Anolis carolinensis;* **Gg-** *Gallus gallus;* **Oa-** *Ornithorhynchus anatinus;* **Tr-** *Takifugu rubripes;* **Ol-** *Oryzias latipes;* **Ga-** *Gasterosteus aculeatus*; **Tn-** *Tetraodon nigroviridis;* **Dr-** *Danio rerio*

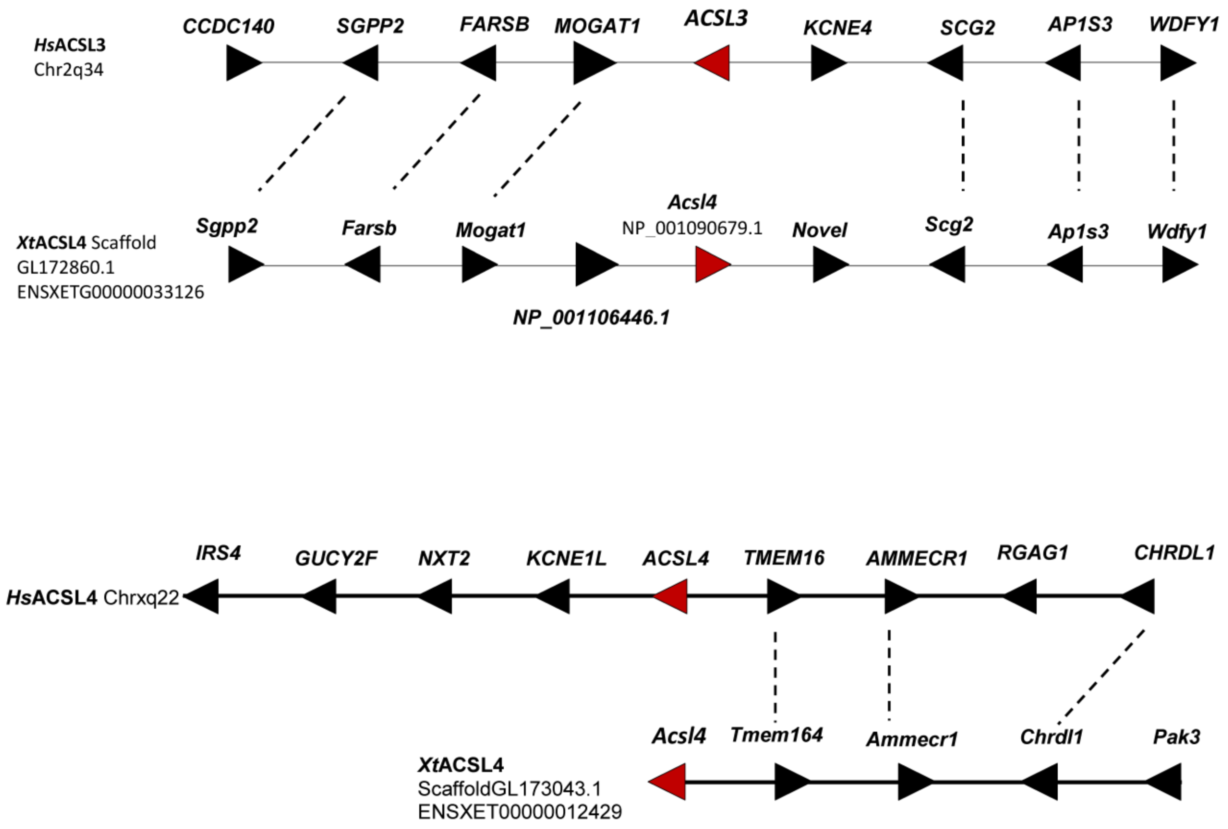
Additional file 4

## Synteny maps of Zebrafish ACSL4  3R duplicates



Additional file 5

## *Xenopus tropicalis Acsl3* and *Acsl4* and corresponding genome location in human



# CHAPTER IV

## FATTY ACID BIOSYNTHESIS

# Chapter IV – Fatty Acid Biosynthesis

## IV.1 Evolutionary functional elaboration of the Elovl2/5 gene family in chordates

Óscar Monroig*, Mónica Lopes-Marques*, Juan C. Navarro, Francisco Hontoria,

Raquel Ruivo, MigueM. Santos, Byrappa Venkatesh, Douglas R. Tocher, L. Filipe C. Castro

(*Joint first authors)

Scientific Reports

(DOI:10.1038/srep20510)

SCIENTIFIC REPORTS

OPEN

# Evolutionary functional elaboration of the *Elovl2/5* gene family in chordates

Received: 09 October 2015
Accepted: 05 January 2016
Published: 09 February 2016

Óscar Monroig[1,*], Mónica Lopes-Marques[2,3,*], Juan C. Navarro[4], Francisco Hontoria[4], Raquel Ruivo[2], Miguel M. Santos[2,5], Byrappa Venkatesh[6], Douglas R. Tocher[1] & L. Filipe C. Castro[2,5]

**The biosynthesis of long-chain polyunsaturated fatty acids (LC-PUFA) provides an intriguing example on how multi-enzymatic cascades evolve. Essential LC-PUFA, such as arachidonic, eicosapentaenoic, and docosahexaenoic acids (DHA), can be acquired from the diet but are also endogenously retailored from $C_{18}$ precursors through consecutive elongations and desaturations catalyzed, respectively, by fatty acyl elongase and desaturase enzymes. The molecular wiring of this enzymatic pathway defines the ability of a species to biosynthesize LC-PUFA. Exactly when and how in animal evolution a functional LC-PUFA pathway emerged is still elusive. Here we examine key components of the LC-PUFA cascade, the *Elovl2*/*Elovl5* elongases, from amphioxus, an invertebrate chordate, the sea lamprey, a representative of agnathans, and the elephant shark, a basal jawed vertebrate. We show that *Elovl2* and *Elovl5* emerged from genome duplications in vertebrate ancestry. The single *Elovl2/5* from amphioxus efficiently elongates $C_{18}$ and $C_{20}$ and, to a marked lesser extent, $C_{22}$ LC-PUFA. Lamprey is incapable of elongating $C_{22}$ substrates. The elephant shark *Elovl2* showed that the ability to efficiently elongate $C_{22}$ PUFA and thus to synthesize DHA through the Sprecher pathway, emerged in the jawed vertebrate ancestor. Our findings illustrate how non-integrated "*metabolic islands*" evolve into fully wired pathways upon duplication and neofunctionalization.**

The origin of complexity in living systems is a central question in evolution[1,2]. Pairwise interactions between molecules (e.g. ligand and receptors; enzymes and their substrates) and the impact of gene duplication on protein function have provided crucial insight into the understanding of physiological diversity[3]. Additionally, the association of different enzymes into single pathways and how these are affected by evolutionary processes is fundamental to reconstruct the history of metabolic gene networks[4,5]. The biosynthesis of long-chain ($C \geq 20$) polyunsaturated fatty acids (LC-PUFA) in animals represents a fascinating example, where phylogenetically unrelated enzymes participate in a metabolic cascade to synthesize vital molecules such as arachidonic acid (ARA, 20:4n-6), eicosapentaenoic acid (EPA, 20:5n-3) and docosahexaenoic acid (DHA, 22:6n-3)[6,7] (Fig. 1A). In addition to dietary input, LC-PUFA are synthesized endogenously from essential $C_{18}$ polyunsaturated fatty acid (PUFA) precursors including linoleic acid (LOA, 18:2n-6) and $\alpha$-linolenic acid (ALA, 18:3n-3) in mammals and teleosts, through a series of consecutive desaturation and elongation reactions[8] (Fig. 1A). How and when this gene pathway has emerged and functionally diversified over time is still obscure. Typically in mammals, the metabolic cascade converting $C_{18}$ PUFA into bioactive $C_{20-22}$ LC-PUFA, such as ARA, EPA and DHA requires the concerted action of distinct $\Delta 5$ and $\Delta 6$ fatty acyl desaturase (FADS) enzymes, as well as that of elongation of long-chain fatty acids (ELOVL) proteins including ELOVL2 and ELOVL5 at specific steps in the pathway[8] (Fig. 1A). Recently, the ability for direct $\Delta 4$ desaturation of 22:5n-3 to 22:6n-3 has been also shown in human

[1]Institute of Aquaculture, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK. [2]CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, U. Porto – University of Porto, Rua dos Bragas 289, 4050-123 Porto, Portugal. [3]ICBAS - Institute of Biomedical Sciences Abel Salazar, U. Porto - University of Porto, Rua de José Viterbo Ferreira 228, 4050-313 Porto, Portugal. [4]Instituto de Acuicultura Torre de la Sal (IATS-CSIC), Ribera de Cabanes 12595, Castellón, Spain. [5]Department of Biology, Faculty of Sciences, U. Porto - University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal. [6]Institute of Molecular and Cell Biology, Agency for Science, Technology and Research, Biopolis, Singapore 138673. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Ó.M. (email: oscar.monroig@stir.ac.uk) or L.F.C.C. (email: filipe.castro@ciimar.up.pt)

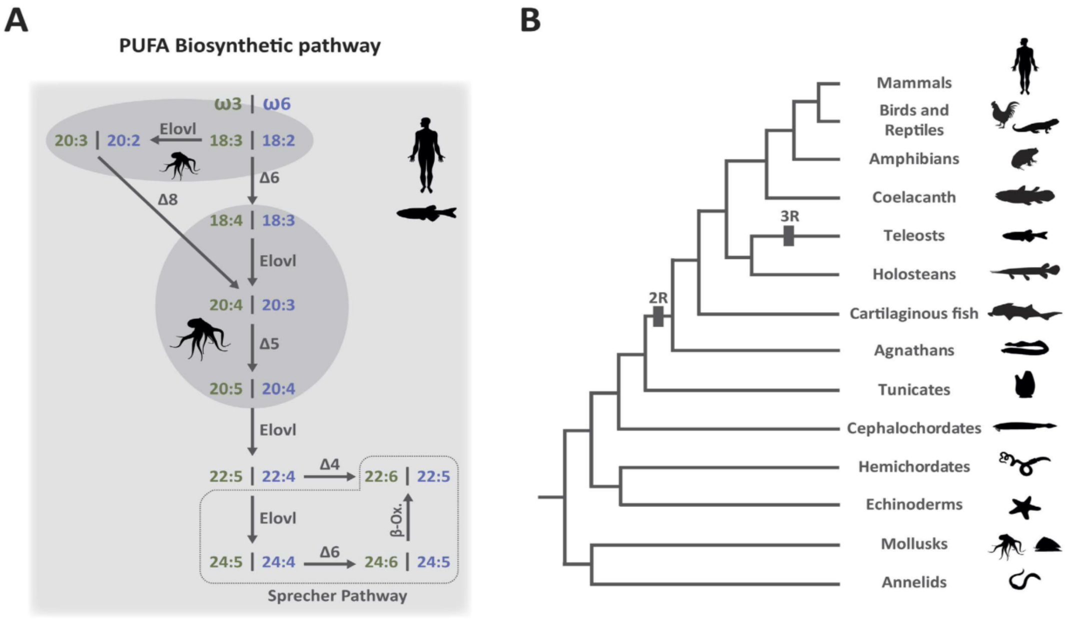


**Figure 1. Biosynthetic pathway of LC-PUFA as determined in mammals and teleosts (all reactions shown), and octopus (confined to reactions in the two ellipses) (A)** . Elongation (Elovl), desaturation (Δ4, Δ5 and Δ6) and β-oxidation (β-oxi) reactions are indicated. The omega-3 (ω3) and omega-6 (ω6) PUFA synthesis cascades are shown in parallel. Each composite number (e.g. 18:3) refers to a specific PUFA, with the first number indicating the number of carbon atoms and the second referring to the ethylenic bonds (details on each PUFA in supplementary Table 1). Phylogenetic tree of the major Bilaterian animal groups considered in this study (**B**). Genome duplications are indicated (2R and 3R).

FADS2[9]. The mechanisms of LC-PUFA biosynthesis in teleost fish, particularly farmed species, have been extensively investigated in the past decades, and many aspects of these metabolic pathways are better understood in fish compared to mammals. For example, the specific ability to convert $C_{18}$ PUFA into LC-PUFA is directly dependent on the exact *Fads* and *Elovl* gene repertoire as well as their substrate specificities[10–13]. It has been shown that the inability of most teleosts to utilize Δ5 desaturase substrates is linked to the specific loss of *Fads1*[11,12]. Surprisingly, the small number of teleost species able to perform Δ5 conversions have a *fads2* gene with Δ5 activity[10,14–16].

Genes encoding ELOVL proteins have received comparatively less attention, although their action is critical for a complete and functional LC-PUFA pathway[17] (Fig. 1A). Generally, mammalian ELOVL5 is involved in the elongation of $C_{18}$ and $C_{20}$ PUFA, whilst ELOVL2 is predominantly active towards $C_{20}$ and $C_{22}$ PUFA[18,19] (Fig. 1A). In contrast, the bird ELOVL5 is, to some extent, able to convert docosapentaenoic acid (DPA, 22:5n-3) to $C_{24}$ LC-PUFA, though with considerable less efficiency than ELOVL2, which displays a similar substrate preference to mammals[20]. The *elovl* gene repertoire in teleosts is also distinctive from that of tetrapods. Most species studied so far have a single *elovl5* gene with the ability to elongate $C_{18}$ and $C_{20}$ PUFA substrates, with marginal activity towards $C_{22}$[21–25], with Atlantic salmon appearing as the sole fish species where two copies of *elovl5* have been characterized[11,26]. In contrast, an *elovl2* orthologue has been identified only in Atlantic salmon[10] (*Salmo salar*), zebrafish[27] (*Danio rerio*) and rainbow trout[28] (*Oncorhynchus mykiss*), and with ray-finned fishes (including most marine species) appearing to lack *elovl2* in their genomes[11]. Similar to their tetrapod counterparts, teleost *elovl2* demonstrated the capacity to elongate DPA and thus contribute to DHA production through the so-called "Sprecher pathway"[29] (Fig. 1A). From the above, *Elovl5* appears to be unique in its capability to elongate $C_{18}$ PUFA substrates and, similarly, *Elovl2* towards $C_{22}$ PUFA, while there is an overlap between both enzymes in their capacity to metabolize $C_{20}$ substrates. However, when exactly *Elovl2* and *Elovl5* genes diverged and their respective functional fatty acid preferences emerged in metazoan evolution is presently unknown. Interestingly, various mollusk species, including the common octopus (*Octopus vulgaris*), the noble scallop (*Chlamys nobilis*) and cuttlefish (*Sepia officinalis*), have been shown to possess an *Elovl* gene, phylogenetically basal to the vertebrate *Elovl2* and *Elovl5*[30–32]. Curiously, the mollusk Elovl enzyme is only capable of metabolizing $C_{18}$ PUFA and to lesser extent $C_{20}$[30–32] (Fig. 1A). The desaturase abilities in mollusks are also markedly different when compared to mammals and teleosts, since only Δ5 desaturases have been described so far[33–35] (Fig. 1A). These results suggest a complex scenario regarding the evolutionary emergence of a complete LC-PUFA biosynthetic pathway.

Despite the significant effort made to clarify the LC-PUFA biosynthetic capabilities in some vertebrate lineages, the presently known complement of *Fads* and *Elovl* genes and their biosynthetic abilities in key evolutionary lineages hampers the precise evolutionary profiling of this pathway. Here we investigate the *Elovl2*/*Elovl5* gene repertoire at a key evolutionary moment: the invertebrate/vertebrate transition (Fig. 1B). By examining three species, including the European amphioxus (*Branchiostoma lanceolatum*, cephalochordate), the sea lamprey

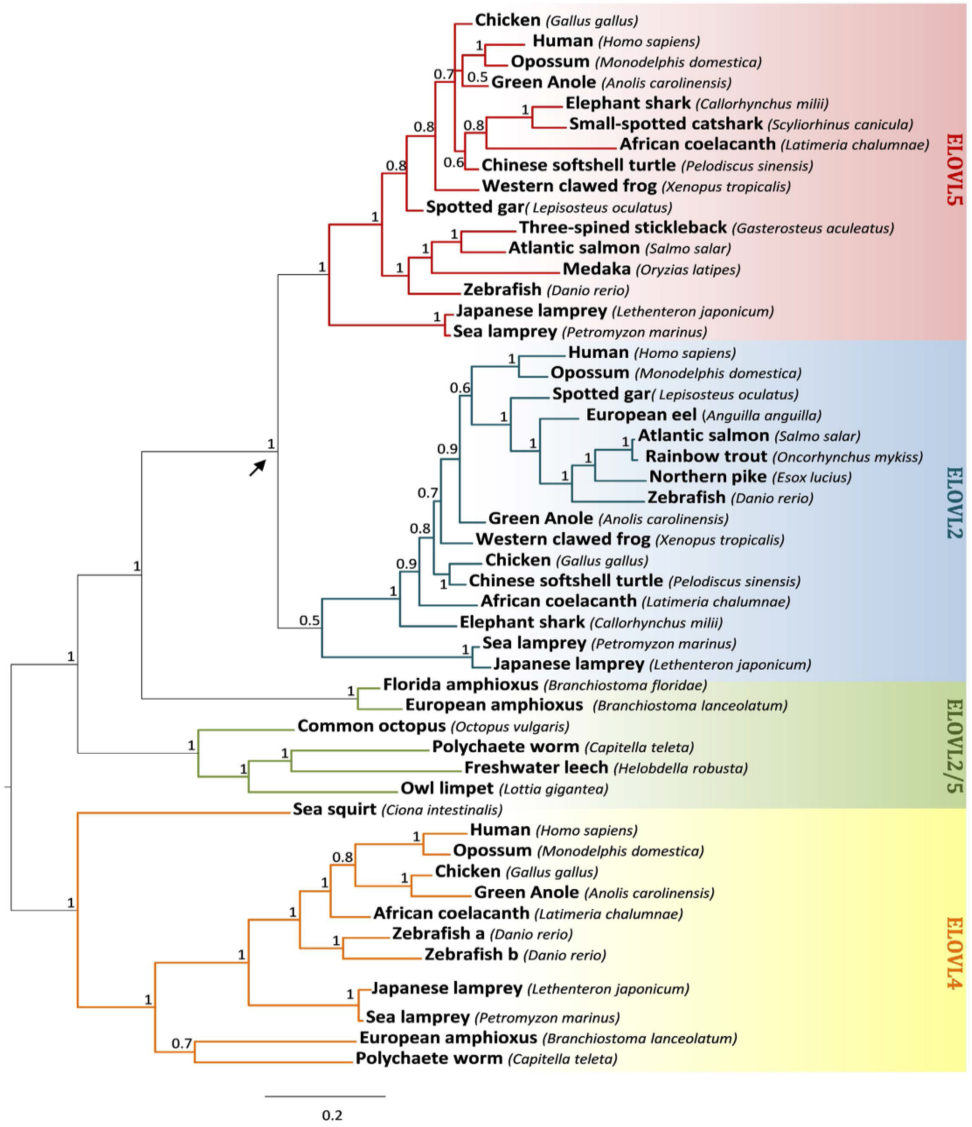

Chicken (Gallus gallus)
Human (Homo sapiens)
Opossum (Monodelphis domestica)
Green Anole (Anolis carolinensis)
Elephant shark (Callorhynchus milii)
Small-spotted catshark (Scyliorhinus canicula)
African coelacanth (Latimeria chalumnae)
Chinese softshell turtle (Pelodiscus sinensis)
Western clawed frog (Xenopus tropicalis)
Spotted gar( Lepisosteus oculatus)
Three-spined stickleback (Gasterosteus aculeatus)
Atlantic salmon (Salmo salar)
Medaka (Oryzias latipes)
Zebrafish (Danio rerio)
Japanese lamprey (Lethenteron japonicum)
Sea lamprey (Petromyzon marinus)
ELOVL5
Human (Homo sapiens)
Opossum (Monodelphis domestica)
Spotted gar( Lepisosteus oculatus)
European eel (Anguilla anguilla)
Atlantic salmon (Salmo salar)
Rainbow trout (Oncorhynchus mykiss)
Northern pike (Esox lucius)
Zebrafish (Danio rerio)
Green Anole (Anolis carolinensis)
Western clawed frog (Xenopus tropicalis)
Chicken (Gallus gallus)
Chinese softshell turtle (Pelodiscus sinensis)
African coelacanth (Latimeria chalumnae)
Elephant shark (Callorhynchus milii)
Sea lamprey (Petromyzon marinus)
Japanese lamprey (Lethenteron japonicum)
ELOVL2
Florida amphioxus (Branchiostoma floridae)
European amphioxus (Branchiostoma lanceolatum)
Common octopus (Octopus vulgaris)
Polychaete worm (Capitella teleta)
Freshwater leech (Helobdella robusta)
Owl limpet (Lottia gigantea)
ELOVL2/5
Sea squirt (Ciona intestinalis)
Human (Homo sapiens)
Opossum (Monodelphis domestica)
Chicken (Gallus gallus)
Green Anole (Anolis carolinensis)
African coelacanth (Latimeria chalumnae)
Zebrafish a (Danio rerio)
Zebrafish b (Danio rerio)
Japanese lamprey (Lethenteron japonicum)
Sea lamprey (Petromyzon marinus)
European amphioxus (Branchiostoma lanceolatum)
Polychaete worm (Capitella teleta)
ELOVL4
0.2

**Figure 2. Bayesian molecular phylogenetic analysis of the *Elovl2, Elovl5 and Elovl4* genes.** Numbers at nodes indicate posterior probabilities. Arrow denotes duplication timing of *Elovl2/5*. Rooted on the *Elovl4* clade. Accession numbers for all sequences are provided in the supplementary Table 2.

(*Petromyzon marinus*, agnathan) and the elephant shark (*Callorhinchus milii*, basal gnathostome), we provide an insightful snapshot into the evolution of critical enzymes dictating the LC-PUFA biosynthetic pathways in chordates.

## Results

**Elovl2 and Elovl5 originated in the ancestor of vertebrates.** We analyzed the repertoire of *Elovl2* and *Elovl5* like genes in a total of 19 species representing all major vertebrate lineages (Sarcopterigii, Actinopterigii, Chondrichthyans and Agnathans) (Fig. 1B). In addition, we also investigated invertebrate species, representing four phyla from invertebrate protostomes and deuterostomes (Fig. 1B). The retrieved sequence dataset was used for phylogenetic reconstruction employing two methods, Bayesian analysis (BA) and Maximum likelihood (ML) (supplementary Fig. 1 for the ML phylogeny). We found two well-supported monophyletic groups, one containing all *Elovl4* sequences, and another containing invertebrate single copy *Elovl2/5* from cephalochordates and various protostome species and all vertebrate *Elovl2* and *Elovl5* sequences (Fig. 2). Within the latter group, gnathostome sequences formed two sister clades *Elovl2* and *Elovl5*, respectively. Each of the lamprey sequences branched together with gnathostome *Elovl2* and *Elovl5*, although with low statistical support in the case of *Elovl2* (Fig. 2). Therefore, the overall tree topology is indicative of the timing of *Elovl2/5* gene expansion, coincident with the evolution of the vertebrate lineage approximately 500 million years ago. No tunicate sequences were used in our analyses since no orthologues of *Elovl2/5* were found in genomes from sea squirts (*Ciona intestinatis* and *C. savignyi*) and the star ascidian (*Botryllus schlosseri*), despite the former having an *Elovl4*-like gene with the ability to elongate $C_{18}$ and $C_{20}$ PUFA[36]. Additionally, while some studies in teleosts have suggested that *Elovl4* can partly contribute to the LC-PUFA biosynthesis[37] these enzymes are generally related to the biosynthesis of very long-chain ($C > 24$) fatty acids[38], and thus were not considered in this study.

**Genome duplications generated Elovl2 and Elovl5 paralogues in vertebrates.** The phylogenetic analysis supported the timing of *Elovl2* and *Elovl5* origin to the ancestor of vertebrates. Thus, we hypothesize that genome duplications were involved in the diversification of *Elovl2/5* genes. Human *Elovl2* and *Elovl5* localize to the same human chromosome (Hsa6) though at separate regions (Fig. 3A). These two genomic sections were linked to a four-fold paralogy originating from genome duplications[38] (linkage group 4) involving a quartet of regions: paralogy A at Hsa20.5, paralogy B at Hsa2.1/6.6/6.8, paralogy C at Hsa6.2/8.2/8.4, and paralogy D at Hsa1.2 (supplementary Fig. 2). In effect, neighboring *Elovl2* and *Elovl5* gene families with a duplication history coincident with genome duplications, have, in most cases, a gene-by-gene paralogy in the expected regions (Fig. 3A). For example, *GCM1* and *ICK* (neighbors of *Elovl5*) have a vertebrate specific paralog mapping to the *Elovl2 locus*, *GCM2* and *MAK*, respectively. Also *Sycp2l*, localizing close to *Elovl2*, has a paralogue at Hsa20.5 as expected (paralogy group A) (Fig. 3A). Additionally, we also examined the *Elovl2/5* genomic *locus* of the pre-duplicated genome of the Florida amphioxus (*B. floridae*) (Fig. 3B). Coherently, we found that the neighboring genes *Bag2*, *Ndufs5* and *Mfsd2* have their human orthologues localizing to Hsa6 (close to *Elovl5*) and Hsa1, part of linkage group C and D, respectively (Fig. 3B, supplementary Fig. 2). Thus, we can conclude that *Elovl2* and *Elovl5* have appeared as part of whole-genome duplications.

**Are Agnathan Elovl genes exact Elovl2 and Elovl5 orthologues?** To further clarify the orthology of the identified *Elovl2/5* sequences, we examined the syntenic relationships of *Elovl2/5* genes in key species. Gnathostome *Elovl2* and *Elovl5* gene *loci* were conserved, though with different degrees (Fig. 3C, D; supplementary Fig. 3). For example, *Sycp2l* flanks *Elovl2* in humans and the elephant shark, indicative of a common origin (Fig. 3C). A strongly conserved syntenic pattern was also observed in the *Elovl5 locus*, with *Gcm1* and *Gclc* outflanking this gene in all gnathostome species except the former in zebrafish (Fig. 3D; supplementary Fig. 3). The exact orthology of agnathan gene sequences poses some challenges, namely when evolutionary processes such as whole genome duplications and gene loss are involved[40–42]. Given that the putative lamprey *Elovl2* was statistically weakly supported in the phylogenetic tree, we examined also the flanking gene families of the putative *Elovl* genes in both the sea lamprey and the Japanese lamprey (*Lethenteron japonicum*). In both species, the putative *Elovl2 locus* includes orthologues of *Sycp2l* and *Gcm2* gene, denoting a strong conservation with the human *locus* (Fig. 3C). In contrast, the "*Elovl5*" *locus* of lampreys displays no synteny conservation with other vertebrates (Fig. 3D). Although we cannot exclude that this represents a different paralogue retained uniquely in lampreys, we suggest that this is a *bona fide Elovl5* gene, in a highly rearranged *locus*.

**Functional characterization of amphioxus, sea lamprey and elephant shark ELOVL enzymes.** We next analyzed the substrate specificities of ELOVL enzymes from three chordate species, namely amphioxus, sea lamprey and elephant shark (Table 1). Transgenic yeast expressing the amphioxus *Elovl2/5* ORF were able to elongate $C_{18}$, $C_{20}$ and, to a lesser extent, $C_{22}$ PUFA substrates (Table 1). The sea lamprey *Elovl5* showed relatively high activity towards $C_{18}$ PUFA (18:4n-3 and 18:3n-6), and lower activity toward the $C_{20}$ PUFA (20:5n-3 and 20:4n-6). Compared to the sea lamprey *Elovl5*, the *Elovl2* was very efficient in the elongation of $C_{20}$ to $C_{22}$, with $C_{18}$ PUFA being elongated to a lesser extent (Table 1). Interestingly, neither of the sea lamprey *Elovl* enzymes displayed the capacity to elongate $C_{22}$ to $C_{24}$ (Table 1). In order to investigate when the *Elovl2* acquired the ability to elongate $C_{22}$ PUFA, we tested the function of the elephant shark *Elovl2*. Consistent with the activities exhibited by fish and mammalian orthologues[16,43] the elephant shark *Elovl2* had marginal activity towards $C_{18}$ PUFA and high elongation capability on $C_{20}$ and $C_{22}$ PUFA that were converted into the corresponding $C_{22}$ and $C_{24}$ elongation products, respectively (Table 1). Moreover, the functional characterization of the elephant shark *Elovl5* confirmed its ability to elongate preferably $C_{18}$ and $C_{20}$ to $C_{20}$ and $C_{22}$ PUFA (Table 1), respectively, as typically observed in other vertebrate lineages[16,17].

**W231C substitution confers $C_{22}$ to $C_{24}$ elongation capacity to sea lamprey Elovl2.** Functional characterization of the sea lamprey *Elovl2* showed no ability to elongate $C_{22}$ PUFA to $C_{24}$ products contrary to those of gnathostome *Elovl2*. On the other hand, elephant shark *Elovl2*, whose sequence contains the specific cysteine (C) residue regarded as critical for elongation of $C_{22}$ by *Elovl2*[41] (supplementary Fig. 4), showed ability to elongate $C_{22}$ PUFA as in gnathostome lineages. Coherently, the sea lamprey *Elovl2* exhibits a tryptophan (W) typical of *Elovl5* sequences (supplementary Fig. 4). Thus, we next tested whether site-directed mutagenesis of W231C would drift the enzymatic activity towards $C_{22}$ PUFA elongation as observed in the gnathostome orthologue. Our mutagenesis analysis showed that the W231C substitution conferred the sea lamprey *Elovl2* the ability to elongate 22:5n-3 to 24:5n-3, although the conversion obtained in the yeast expression system (2%) was notably lower when compared to other *Elovl2* proteins characterized in the present study and previously reported using similar systems[11,27] (Table 1). Interestingly, the mutant retained its ability to elongate $C_{20}$ PUFA such as 20:5n-3 and 20:4n-6 to the corresponding $C_{22}$ PUFA, 22:5n-3 and 22:4n-6, but lost its ability to elongate $C_{18}$ PUFA (Table 1). Overall, the functional characterization of sea lamprey *Elovl2* mutant confirms that the cysteine (C) residue indicated above is key for the $C_{22}$ to $C_{24}$ elongation ability[42], but the relatively low conversion observed in the yeast system suggests that other amino acids are also critical for an efficient conversion of $C_{22}$ into $C_{24}$ PUFA.

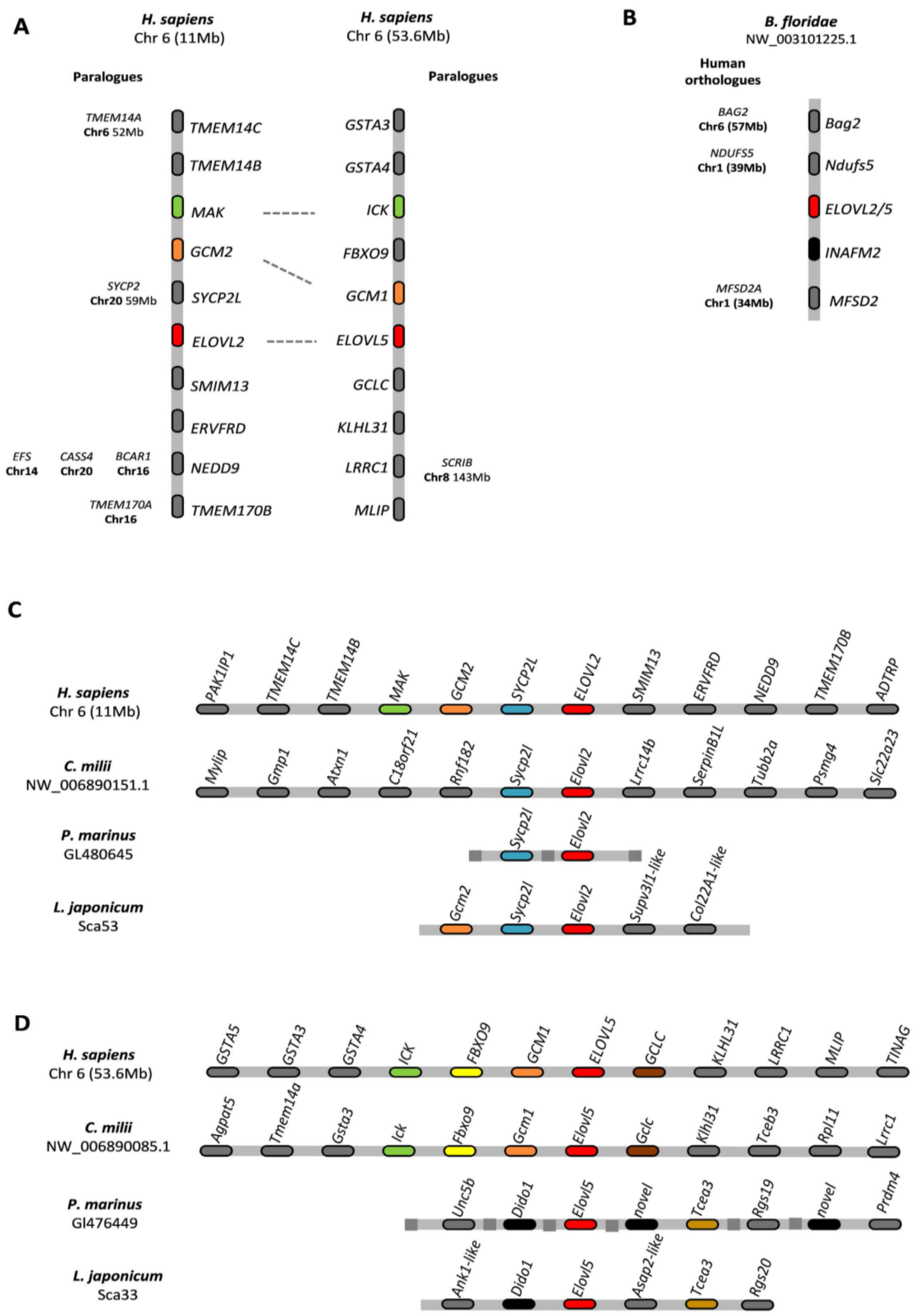


**Figure 3. Comparative genomic maps of *Elovl* gene *loci*.** (**A**) Paralogy analysis of *ELOVL2* and *ELOVL5* human orthologues; (**B**) the amphioxus *elovl2/5* gene *locus*; (**C**) synteny analysis of the *Elovl2* genes in lampreys, human and elephant shark; (**D**) synteny analysis of the *Elovl2* genes in lampreys, human and elephant shark.

## Discussion

Vertebrate radiation encompassed the acquisition of key physiological and anatomical innovations, as a consequence of gene and genome duplications[44–47]. Among others, these might have facilitated the challenge of colonizing new ecological niches with diverse nutrient composition such as, for example, LC-PUFA. ELOVL are key enzymes involved in the rate-limiting step of fatty acid elongation pathway by which β-ketoacyl-CoA is produced after the condensation of acyl-CoA molecule and malonyl-CoA[17]. Although these enzymes have been

| FA substrate | FA product | *Amphioxus Elovl2/5* | *Sea lamprey Elovl5* | *Sea lamprey Elovl2* | *Sea lamprey mutated Elovl2* | *Elephant shark Elovl5* | *Elephant shark Elovl2* |
|---|---|---|---|---|---|---|---|
| 18:4n-3 | 20:4n-3 | 21 | 56 | 9 | 0 | 69 | 7 |
| 18:3n-6 | 20:3n-6 | 55 | 40 | 0 | 0 | 74 | 3 |
| 20:5n-3 | 22:5n-3 | 87 | 12 | 88 | 57 | 65 | 85 |
| 20:4n-6 | 22:4n-6 | 88 | 8 | 25 | 8 | 56 | 82 |
| 22:5n-3 | 24:5n-3 | 14 | 0 | 0 | 2 | 5 | 43 |
| 22:4n-6 | 24:4n-6 | 4 | 0 | 0 | 0 | 2 | 37 |

**Table 1. Functional characterization of the amphioxus *Elovl2/5*, the sea lamprey *Elovl5*, *Elovl2* and mutated *Elovl2*, and the elephant shark *Elovl5* and *Elovl2* in *Saccharomyces cerevisiae*.** Conversions were calculated according to the formula (all product areas/(all products areas + substrate area)) × 10.

extensively studied in a number of metazoans including invertebrates and vertebrates, their evolution has yet to be deciphered. Here we focused on a subset of *Elovl* genes, namely *Elovl2* and *Elovl5*, critical in the biosynthetic pathways of LC-PUFA[17]. Combining phylogenetics, comparative genomics and functional data, we have been able to deduce the early evolution of functional *Elovl* specificities in chordates.

Phylogenetics and synteny revealed that orthologs of *Elovl2* and *Elovl5* occur only in vertebrate species. Thus, our data support that both *Elovl5* and *Elovl2* have evolved in agnathans, chondrichthyans, holosteans (spotted gar) and teleosts such as zebrafish and Atlantic salmon. Contrarily to *Elovl5*, *Elovl2* is absent in most of the ray-finned fish branch due to a gene loss event as previously hypothesized[11]. Moreover, the finding of a single *Elovl2/5* sequence in invertebrate deuterostomes and protostomes, and its basal position in the tree, defines the transition from invertebrate chordates to vertebrates as the exact timing at which diversification of *Elovl2/5* gene family occurred. Typically in mammals, ELOVL2 are enzymes with high elongation efficiency towards $C_{20}$ and $C_{22}$ PUFA, and marginal (if any) activity towards $C_{18}$ substrates[8,17] (Fig. 4A). In contrast, ELOVL5 have $C_{18}$ and $C_{20}$ PUFA as preferred substrates, but have little or no capability to elongate $C_{22}$ PUFA[8,17]. The former elongation specificity is largely exhibited by protostomes such as the octopus *Elovl2/5*[30] (Fig. 4A). In agreement, the here reported amphioxus ELOVL2/5 enzyme showed the same elongation pattern with $C_{18}$ and $C_{20}$ PUFA appearing as preferred substrates for elongation, although some ability to elongate $C_{22}$ PUFA was also observed (Fig. 4A). The substrate preferences of the sea lamprey *Elovl2* and the *Elovl5* enzymes showed a complete inability to elongate $C_{22}$ PUFA, whereas the elephant shark *Elovl2* was able to effectively elongate $C_{22}$ PUFA, 22:5n-3 and 22:4n-6, to their corresponding $C_{24}$ products as shown in teleosts and mammalian ELOVL2 proteins[11,17,27,28] (Fig. 4A). The amino acid alignment of the various *Elovl2/5* sequences allowed us to identify that the elephant shark *Elovl2*, similar to orthologues from other gnathostome lineages including mammals, birds, amphibians and teleosts, contained within its sequence the cysteine (C) regarded as critical for $C_{22}$ PUFA elongation by *Elovl2*[43], while this residue was substituted by a tryptophan (W) in the sea lamprey *Elovl2*. Using a site-directed mutagenesis approach, we showed that the mutated lamprey *Elovl2* protein lost the ability to elongate $C_{18}$ PUFA exhibited by the native protein and, more importantly, gained the ability to elongate $C_{22}$ PUFA. However, the minute capacity to elongate $C_{22}$ exhibited by the mutated lamprey *Elovl2* suggests that other unidentified amino acids are also critical for this function.

Apart from elongase activity, the complexity of the LC-PUFA biosynthetic network cannot be dissociated from LC-PUFA desaturation profiles. The combined analysis of *Fads* and *Elovl* gene repertoire and function in various species allows us to propose that a fragmented LC-PUFA pathway existed early in evolution (Fig. 4B). Data derived from mollusks strongly suggests that the ancestral bilaterian LC-PUFA biosynthetic pathway was composed of *Fads* and *Elovl* genes encoding, respectively, proteins with single desaturation ($\Delta 5$) and elongation ($C_{18}$ to $C_{22}$) enzymatic abilities[30–35] (Fig. 4B), although the presence of additional uncharacterized desaturases in mollusks impedes a final conclusion[48]. An incomplete pathway also appears to exist in cephalochordates. Despite the functionalities of *Elovl2/5* showing its ability to elongate PUFA ranging from $C_{18}$ to $C_{22}$, a full complement of desaturase abilities is likely absent as suggested by *in silico* searches, with a single *Fads*-like gene described so far in their genome[12] (Fig. 4B). However, relevant levels of DHA were found in the digestive tract of amphioxus[49]. While they could be exclusively diet-derived, an endogenous DHA production cannot be excluded. Thus, the characterization of the single amphioxus FADS should be addressed in the future. In agnathans, on the other hand, the restricted elongation profiles demonstrated by the lack of elongation activity by both Elovl-like enzymes towards $C_{22}$ may limit the LC-PUFA biosynthetic pathways regardless of the possible number of genes or desaturase activities existing in lampreys (Fig. 4B). Importantly, the combined activities of the elephant shark *Elovl5* and *Elovl2* enabling elongation up to $C_{24}$ LC-PUFA and thus DHA biosynthesis[29], as well as the existence of $\Delta 6$ and $\Delta 5$ Fads in chondrichthyans[12], strongly suggest that a fully developed LC-PUFA biosynthetic pathway dependent on the sequential action of *Elovl* and *Fads* was first operational in gnathostomes (Fig. 4B). The overall pathway has been conserved throughout this lineage with localized episodes of gene loss, gene duplication and functional plasticity as demonstrated by the $\Delta 5$ capacity of some teleost *Fads2*[12].

However, it is difficult to foresee the exact evolutionary drivers accounting for the acquisition of a full biosynthetic pathway for LC- PUFA in organisms that have a likely supply in the diet. Clearly though, endogenous production of DHA, the final LC-PUFA in the cascade, is physiologically advantageous since it represents an additional source to cope with potential dietary scarcity, as well as satisfy particularly high requirements in early development[50]. Additionally, DHA levels are known to be especially high in tissues such as brain and retina, in mammals, teleosts and chondrichthyans[27,50,51]. Thus, it is conceivable to hypothesize that the elaboration of brain

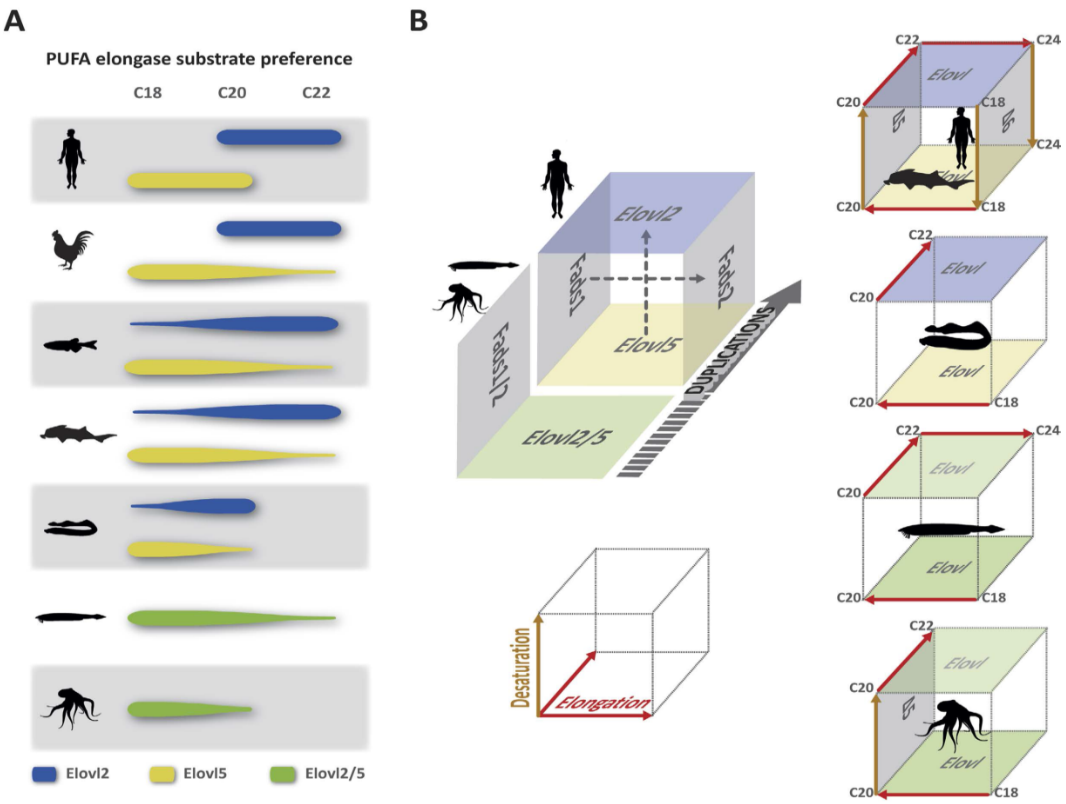


**Figure 4. Evolutionary scenario of LC-PUFA biosynthesis functional diversification in Bilateria.** (**A**) PUFA elongase substrate preference in octopus, amphioxus, lamprey, elephant shark, zebrafish, chicken and human; (**B**) schematic view of gene duplication events in the elongation/desaturation network along the invertebrate/vertebrate transition (top left) and known enzymatic activities of FADS and ELOVL (right; from top to bottom: octopus, amphioxus, lamprey and the gnathostomes elephant shark and human); Δ5 and Δ6 denote desaturation activities.

and eye function in vertebrate ancestry[52], was paralleled by the capacity to endogenously regulate and synthesize DHA independently of exogenous sources.

In conclusion, the observed lineage-specific LC-PUFA biosynthetic profiles in chordate species were tailored by gene duplication events followed by enzymatic neofunctionalizations. We propose that the biosynthesis of the essential fatty acid DHA through the Sprecher pathway from $C_{18}$ precursors was not fully resolved until gnathostomes emerged.

## Methods

**Sequence collection.** ELOVL amino acid (aa) sequences were retrieved from Ensembl, GenBank, JGI (Joint Genome Institute), elephant shark genome project (http://esharkgenome.imcb.a-star.edu.sg/) and Japanese lamprey genome project (http://jlampreygenome.imcb.a-star.edu.sg/), databases through Blastp searches using as reference the annotated human ELOVL2, ELOVL5 and ELOVL4 aa sequences. Accession numbers are available in supplementary Table 2.

**Phylogenetic analysis.** A total of 50 ELOVL aa sequences were aligned with MAFFT[53] (L-INS-i method). The sequence alignment was stripped from all columns containing gaps leaving 200 gap free sites for phylogenetic analysis. Bayesian phylogenetic analysis was performed using MrBayes v3.2.3 available in CIPRES Science Gateway V3.3[54]. MrBayes was run for 5 million generations with the following parameters: rate matrix for aa = mixed, nruns = 2, nchains = 4, temp = 0.2, sampling set to 500 and burin to 0.25. Maximum likelihood phylogenetic analysis was performed in PhyML v3.0 server[55] protein evolutionary model was calculated in PhyML using smart model selection resulting in JTT +G6 +I +F and the number of bootstrap replicates was set to 1000. The resulting trees were visualized in Fig Tree V1.3.1 available at http://tree.bio.ed.ac.uk/software/figtree/ and rooted with ELOVL4 sequences.

**Synteny and comparative genomics.** *Elovl*2 and *Elovl*5 genes were mapped onto the respective species genomes, using the latest genome assemblies available in Ensembl release (Ensembl release 80, May 2015). The elephant shark genomic information was collected from Ensembl Pre assembly ESHARK1 (http://ensembl.

fugu-sg.org/index.html) and for Japanese lamprey synteny maps were inferred using the draft assembly LetJap1.0 available at GenBank. When possible, we analyzed a 1Mb window centered on the corresponding *Elovl* gene, using the human *locus* as reference for comparison. Paralogy studies used the ancestral chordate genome reconstruction[39]. Ensembl paralog and ortholog prediction tools were used to infer evolutionary history of flanking *Elovl* genes in addition to phylogenetic analysis reconfirmation using ML methods.

**Elovl full ORF genes in amphioxus, sea lamprey and elephant shark.** Total RNA was isolated from amphioxus (whole animal) and sea lamprey (kidney, liver and brain) using an Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, UK). All steps were performed according to the manufacturer's recommendations, including the on-column treatment of isolated RNA with RNase-free DNaseI. One μg RNA was used for cDNA synthesis using the iScript cDNA Synthesis Kit (Bio-Rad) and following the manufacturer's specifications. Initial isolation of the *Elovl-like* gene in amphioxus was achieved by PCR with Phusion® Flash (high-fidelity PCR master mix) using degenerate primers (supplementary Table 3). Initial PCR product was confirmed by sequencing and used to design gene specific primers (GSP) used obtain the full-length cDNA sequences by RACE PCR (SMARTer™ RACE cDNA Amplification, Clontech). For the sea lamprey, one complete and one incomplete *Elovl2/5-like* sequences were identified in the available genome. To obtain the open reading frames (ORF) of the incomplete *Elovl2/5-like* sequence (ENSPMAG00000005149), we carried out a RACE PCR. The elephant shark *Elovl2* sequence was identified in the transcriptome and genome sequence[56], and was chemically synthesized (Integrated DNA Technologies, Inc., Glasgow, UK).

**Cloning into pYES2 vector and functional assays in yeast.** Functional characterization of the ELOVL gene products from amphioxus, sea lamprey and elephant shark were investigated by heterologous expression in yeast *Saccharomyces cerevisiae* (strain InvSc1, Invitrogen). Briefly, the ORF of the target genes were cloned into the yeast expression vector pYES2 (Invitrogen) following a two-step routine. First, PCRs with specific primers flanking the full ORF were designed in the 5′ and 3′ UTR of each gene (supplementary Table 3) were performed using Phusion® Flash (high-fidelity PCR master mix) under the following conditions: initial denaturation at 98 °C for 10 s, followed by 25 cycles at 98 °C for 1 s annealing for 5 s and 72 °C for the required amount of time according to the product size. The second step consisted in re-amplification of the initial PCR product (diluted 1/50) with a set of primers containing the start and stop codons and restriction enzyme sites for further cloning into pYES2 (supplementary Table 3). PCR conditions were the same with the exception of the number of cycles that was increased to 35. The resulting PCR product was purified, digested with appropriate restriction enzymes and ligated into a similarly restricted pYES2 vector to produce the constructs pYES2-BlELOVL for *B. lanceolatum Elovl2/5*, pYES2-PmELOVL2 and pYES2-PmELOVL5 for *P. marinus Elovl2* and *Elovl5*, respectively, and pYES2-CmELOVL2 and pYES2-CmELOVL5 for *C. milii Elovl2* and *Elovl5*, respectively. Lamprey *Elovl2* W231C mutant was produced by site directed mutagenesis PCR using pYES2-PmELOVL2 as template, and the PCR product was subsequently purified, digested with the restriction enzymes and ligated into pYES2 to produce pYES2-PmELOVL2-W231C. Accuracy of the DNA sequences was confirmed in all constructs by sequencing. Transformation and culture of yeast were conducted as previously described[10,21]. In order to assess the substrate specificity of the ELOVL enzymes from amphioxus, sea lamprey and elephant shark, transgenic yeast expressing the *Elovl* ORF were grown in the presence of the following PUFA substrates: 18:4n-3, 18:3n-6, 20:5n-3, 20:4n-6, 22:5n-3 and 22:4n-6. After 48 h of incubation, yeast were harvested, washed and total lipid extracted by homogenization in chloroform/methanol (2:1, v/v) containing 0.01% BHT[13].

**Fatty acid analysis of yeast and elongation conversions.** Fatty acyl methyl esters (FAME), prepared from total lipids extracted from harvested cells, were analyzed using a Thermo Gas Chromatograph (Thermo Trace GC Ultra, Thermo Electron Corporation, Waltham, MA, USA) fitted with an on-column injection system and a FID detector. Further confirmation of FAME was performed with an Agilent 6850 Gas Chromatograph system coupled to a 5975 series MSD (Agilent Technologies, Santa Clara, CA, USA). The elongation conversion efficiencies from exogenously added PUFA substrates were calculated by the proportion of substrate fatty acid converted to elongated products as (all product areas/(all product areas + substrate area)) x 100.

## References

1. Lynch, M. & Conery, J. S. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290,** 1151–1155 (2000).
2. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat. Rev. Genet.* **8,** 675–688 (2007).
3. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science* **312,** 97–101 (2006).
4. Phillips, P. C. Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9,** 855–867 (2008).
5. Guo, Z., Jiang, W., Lages, N., Borcherds, W. & Wang, D. Relationship between gene duplicability and diversifiability in the topology of biochemical networks. *BMC Genomics* **15,** 577 (2014).
6. Tocher, D. R. Metabolism and Functions of Lipids and Fatty Acids in Teleost Fish. *Rev. Fish. Sci.* **11,** 107–184 (2003).
7. Schmitz, G. & Ecker, J. The opposing effects of n−3 and n−6 fatty acids. *Prog. Lipid Res.* **47,** 147–155 (2008).
8. Guillou, H., Zadravec, D., Martin, P. G. & Jacobsson, A. The key roles of elongases and desaturases in mammalian fatty acid metabolism: Insights from transgenic mice. *Prog. Lipid. Res.* **49,** 186–199 (2010).
9. Park, H. G., Park, W. J., Kothapalli, K. S. & Brenna J. T. The fatty acid desaturase 2 (FADS2) gene product catalyzes Δ4 desaturation to yield n-3 docosahexaenoic acid and n-6 docosapentaenoic acid in human cells. *FASEB J.* **29,** 3911–9 (2015).
10. Hastings, N. *et al.* A vertebrate fatty acid desaturase with Δ5 and Δ6 activities. *Proc. Natl. Acad. Sci. USA* **98,** 14304–14309 (2001).
11. Morais, S., Monroig, O., Zheng, X., Leaver, M. J. & Tocher, D. R. Highly unsaturated fatty acid synthesis in Atlantic salmon: characterization of ELOVL5- and ELOVL2-like elongases. *Mar. Biotechnol. (NY)* **11,** 627–639 (2009).

12. Castro, L. F. C. *et al.* Functional Desaturase Fads1 (Δ5) and Fads2 (Δ6) Orthologues Evolved before the Origin of Jawed Vertebrates. *PLoS ONE* **7,** e31950 (2012).
13. Monroig, Ó., Tocher, D. R., Hontoria, F. & Navarro, J. C. Functional characterisation of a Fads2 fatty acyl desaturase with Δ6/Δ8 activity and an Elovl5 with C16, C18 and C20 elongase activity in the anadromous teleost meagre (Argyrosomus regius). *Aquaculture.* **412–413,** 14–22 (2013).
14. Li, Y. *et al.* Vertebrate fatty acyl desaturase with Δ4 activity. *Proc. Natl. Acad. Sci. USA* **107,** 16840–16845 (2010).
15. Tanomman, S., Ketudat-Cairns, M., Jangprai, A. & Boonanuntanasarn, S. Characterization of fatty acid delta-6 desaturase gene in Nile tilapia and heterogenous expression in Saccharomyces cerevisiae. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **166,** 148–156 (2013).
16. Fonseca-Madrigal, J. *et al.* Diversification of substrate specificities in teleostei Fads2: characterization of Δ4 and Δ6Δ5 desaturases of Chirostoma estor. *J. Lipid Res.* **55,** 1408–1419 (2014).
17. Jakobsson, A., Westerberg, R. & Jacobsson, A. Fatty acid elongases in mammals: their regulation and roles in metabolism. *Prog. Lipid Res.* **45,** 237–249 (2006).
18. Leonard, A. E. *et al.* Identification and expression of mammalian long-chain PUFA elongation enzymes. *Lipids* **37,733**–740 (2002).
19. Leonard, A. E., Pereira, S. L., Sprecher, H. & Huang, Y. S. Elongation of long-chain fatty acids. *Prog. Lipid Res.* **43,** 36–54 (2004).
20. Gregory, M. K., Geier, M. S., Gibson, R. A. & James, M. J. Functional Characterization of the Chicken Fatty Acid Elongases. *J. Nutr.* **143,** 12–16 (2013).
21. Agaba, M., Tocher, D. R., Dickson, C. A., Dick, J. R. & Teale, A. J. Zebrafish cDNA encoding multifunctional Fatty Acid elongase involved in production of eicosapentaenoic (20:5n-3) and docosahexaenoic (22:6n-3) acids. *Mar. Biotechnol.* (NY) **6,** 251–261 (2004).
22. Zheng, X. *et al.* Physiological roles of fatty acyl desaturases and elongases in marine fish: Characterisation of cDNAs of fatty acyl Δ6 desaturase and elovl5 elongase of cobia (Rachycentron canadum). *Aquaculture* **290,** 122–131 (2009).
23. Gregory, M. K., See, V. H., Gibson, R. A. & Schuller, K. A. Cloning and functional characterisation of a fatty acyl elongase from southern bluefin tuna (Thunnus maccoyii). *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **155,** 178–185 (2010).
24. Mohd-Yusof, N. Y., Monroig, O., Mohd-Adnan, A., Wan, K. L. & Tocher, D. R. Investigation of highly unsaturated fatty acid metabolism in the Asian sea bass, Lates calcarifer. *Fish Physiol. Biochem.* **36,** 827–843 (2010).
25. Morais, S., Mourente, G., Ortega, A., Tocher, J. A. & Tocher, D. R. Expression of fatty acyl desaturase and elongase genes, and evolution of DHA:EPA ratio during development of unfed larvae of Atlantic bluefin tuna (Thunnus thynnus L.). *Aquaculture* **313,** 129–139 (2011).
26. Carmona-Antonanzas, G., Tocher, D. R. & Taggart, J. B. & Leaver, M. J. An evolutionary perspective on Elovl5 fatty acid elongase: comparison of Northern pike and duplicated paralogs from Atlantic salmon. *BMC Evol. Biol.* **13,** 85 (2013).
27. Monroig, Ó., Rotllant, J., Sanchez, E., Cerda-Reverter, J. M. & Tocher, D. R. Expression of long-chain polyunsaturated fatty acid (LC-PUFA) biosynthesis genes during zebrafish Danio rerio early embryogenesis. *Biochim. Biophys. Acta* **1791,** 1093–1101 (2009).
28. Gregory, M. K. & James, M. J. Rainbow trout (Oncorhynchus mykiss) Elovl5 and Elovl2 differ in selectivity for elongation of omega-3 docosapentaenoic acid. *Biochim. Biophys. Acta.* **1841,** 1656–1660 (2014).
29. Sprecher, H. Metabolism of highly unsaturated n-3 and n-6 fatty acids. *Biochim. Biophys. Acta* **1486,** 219–231 (2000).
30. Monroig, Ó., Guinot, D., Hontoria, F., Tocher, D. R. & Navarro, J. C. Biosynthesis of essential fatty acids in Octopus vulgaris (Cuvier, 1797): Molecular cloning, functional characterisation and tissue distribution of a fatty acyl elongase. *Aquaculture* **360–361,** 45–53 (2012).
31. Liu, H. *et al.* Cloning and functional characterization of a polyunsaturated fatty acid elongase in a marine bivalve noble scallop Chlamys nobilis Reeve. *Aquaculture* **416–417,** 146–151 (2013).
32. Monroig, Ó., Hontoria, F., Varó, I., Tocher, D. R. & Navarro, J. C. Investigating the essential fatty acids in the common cuttlefish Sepia officinalis (Mollusca, Cephalopoda): Molecular cloning and functional characterisation of fatty acyl desaturase and elongase. *Aquaculture* **450,** 38–47 (2016).
33. Monroig, Ó., Navarro, J. C., Dick, J. R., Alemany, F. & Tocher, D. R. Identification of a Delta5-like fatty acyl desaturase from the cephalopod Octopus vulgaris (Cuvier 1797) involved in the biosynthesis of essential fatty acids. *Mar. Biotechnol. (NY)* **14,** 411–422 (2012).
34. Li, M. *et al.* Characterization of two Δ5 fatty acyl desaturases in abalone (Haliotis discus hannai Ino). *Aquaculture* **416–417,** 48–56 (2013).
35. Liu, H. *et al.* Functional characterization of a Δ5-like fatty acyl desaturase and its expression during early embryogenesis in the noble scallop Chlamys nobilis Reeve. *Mol. Biol. Rep.* **41,** 7437–7445 (2014).
36. Meyer, A. *et al.* Novel fatty acid elongases and their use for the reconstitution of docosahexaenoic acid biosynthesis. *J. Lipid Res.* **45,** 1899–1909 (2004).
37. Monroig, Ó. *et al.* Expression and role of Elovl4 elongases in biosynthesis of very long-chain fatty acids during zebrafish Danio rerio early embryonic development. *Biochim. Biophys. Acta* **1801,** 1145-1154 (2010).
38. Agbaga, M.-P. *et al.* Role of Stargardt-3 macular dystrophy protein (ELOVL4) in the biosynthesis of very long chain fatty acids. *Proc. Natl. Acad. Sci. USA* **105,** 12843–12848 (2008).
39. Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453,** 1064–1071 (2008).
40. Kuraku, S. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. *Semin. Cell Dev. Biol.* **24,** 119–127 (2013).
41. Mehta, T. K. *et al.* Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). *Proc. Natl. Acad. Sci. USA* **110,** 16044–16049 (2013).
42. Smith, J. J. & Keinath, M. C. The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Res.* **25,** 1081–1090 (2015).
43. Gregory, M. K., Cleland, L. G. & James, M. J. Molecular basis for differential elongation of omega-3 docosapentaenoic acid by the rat Elovl5 and Elovl2. *J. Lipid Res.* **54,** 2851–2857 (2013).
44. Ohno, S. *In Evolution by Gene Duplication* 1st edn,(eds Ohno, S.) Ch. 16, 98–105 (Springer Verlag, 1970).
45. Braasch, I., Volff, J.-N. & Schartl, M. The Endothelin System: Evolution of Vertebrate-Specific Ligand–Receptor Interactions by Three Rounds of Genome Duplication. *Mol. Biol. Evol.* **26,** 783–799 (2009).
46. Minguillon, C., Gibson-Brown, J. J. & Logan, M. P. Tbx4/5 gene duplication and the origin of vertebrate paired appendages. *Proc. Natl. Acad. Sci. USA* **106,** 21726–21730 (2009).
47. Hoffmann, F. G., Opazo J. C. & Storz, J. F. Whole-Genome Duplications Spurred the Functional Diversification of the Globin Gene Superfamily in Vertebrates. *Mol. Biol. Evol.* **29,** 303–312 (2012).
48. Surm, J. M., Prentis, P. J. & Pavasovic, A. Comparative Analysis and Distribution of Omega-3 lcPUFA Biosynthesis Genes in Marine Molluscs. *PLoS One* **10,** e0136301 (2016).
49. Yuan, D. *et al.* Ancestral genetic complexity of arachidonic acid metabolism in Metazoa. *Biochim. Biophys. Acta* **1841,** 1272–84 (2014).
50. Lauritzen, L., Hansen, H. S., Jørgensen, M. H. & Michaelsen, K. F. The essentiality of long chain n-3 fatty acids in relation to development and function of the brain and retina. *Prog. Lipid Res.* **40,**1–94 (2001).
51. Stoknes, I. S., Økland, H. M. W., Falch, E. & Synnes, M. Fatty acid and lipid class composition in eyes and brain from teleosts and elasmobranchs. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **138,** 183–191 (2004).
52. Shimeld, S. M. & Holland, P. W. H. Vertebrate innovations. *Proc. Natl. Acad. Sci. USA* **97,** 4449–4452 (2000).

53. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9,** 286–298 (2008).
54. Miller, M. A. *et al.* A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. *Evol. Bioinform. Online* **11,** 43–48 (2015).
55. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
56. Venkatesh, B. *et al.* Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505,** 174–179 (2014).

## Acknowledgements

This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) (Strategic Funding UID/Multi/04423/2013, SFRH/BD/84238/2012 to M.L.-M. and SFRH/BPD/72519/2010 to R.R.) and partly by MINECO (Spanish Government - AGL2013-40986R). The access to the Institute of Aquaculture laboratories was funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262336 (AQUAEXCEL), Transnational Access Project Number 0095/06/03/13.

## Author Contributions

Ó.M. and L.F.C.C. designed research; Ó.M., M.L.-M., J.C.N., F.H., D.R.T. and L.F.C.C. performed research; Ó.M., M.L.-M., J.C.N., F.H., R.R., M.M.S., B.V., D.R.T. and L.F.C.C. analyzed data; and Ó.M. and L.F.C.C. wrote the paper.

## Additional Information

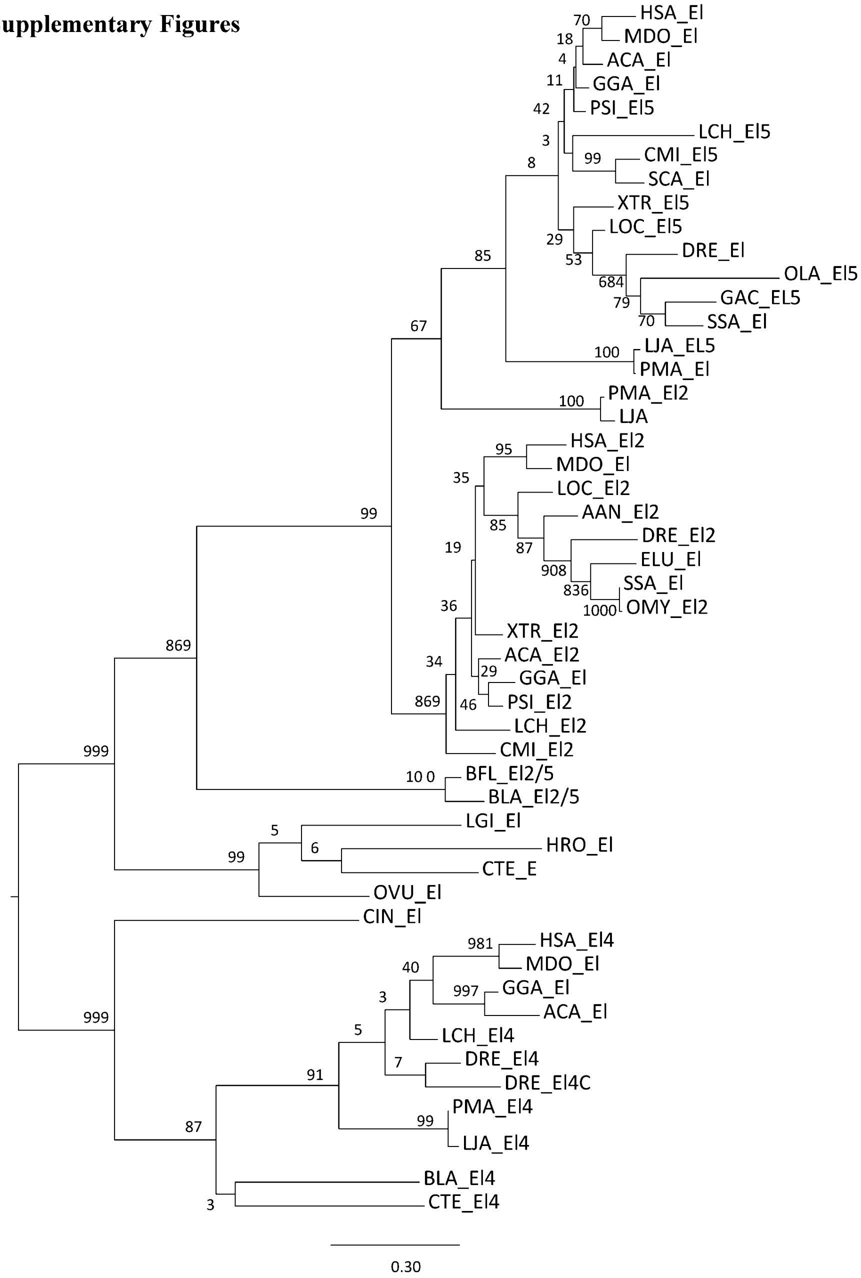
**Supplementary information** accompanies this paper at http://www.nature.com/srep

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article**: Monroig, Óscar. *et al.* Evolutionary functional elaboration of the *Elovl2/5* gene family in chordates. *Sci. Rep.* **6**, 20510; doi: 10.1038/srep20510 (2016).

## SUPPLEMENTARY MATERIAL

### Supplementary Figures



1

**Figure 1.** Maximum likelihood phylogenetic analysis was performed in PhyML v3.0 server (Guindon S et al 2010), protein evolutionary model was calculated in PhyML using smart model selection resulting in JTT +G6 +I +F and the number of bootstrap replicates was set to 1000. HSA – *H. sapiens*; MDO – *M. domestica*; ACA – *A. carolinensis* ; GGA – *G. gallus;* XTR – *X. tropicalis* ; PSI – *P. sinensis;* LCH – *L. chalumnae* ; DRE – *D. rerio;* GAC – *G. aculeatus* ; OMY – *O. mykiss*; OLA – *O. latipes*; SSA – *S. salar;* ELU – *E. lucius;* AAN – *A. anguilla;* CMI – *C. milii;* SCA – *S. canicula;* LJA – *L. japonicum;* PMA – *P. marinus;* LGI - *L. gigantea;* HRO – *H. robusta;* BFL – *B. floridae;* BLA – *B. lanceolatum;* CTE – *C. teleta;* OVU – *O. vulagris;* CIN – *C. intestinalis.*

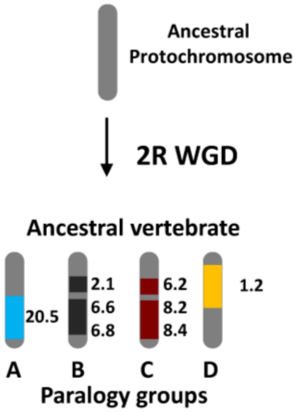


**Figure 2.** Distribution of the ancestral vertebrate paralogy groups containing *Elovl* genes in the human genome.

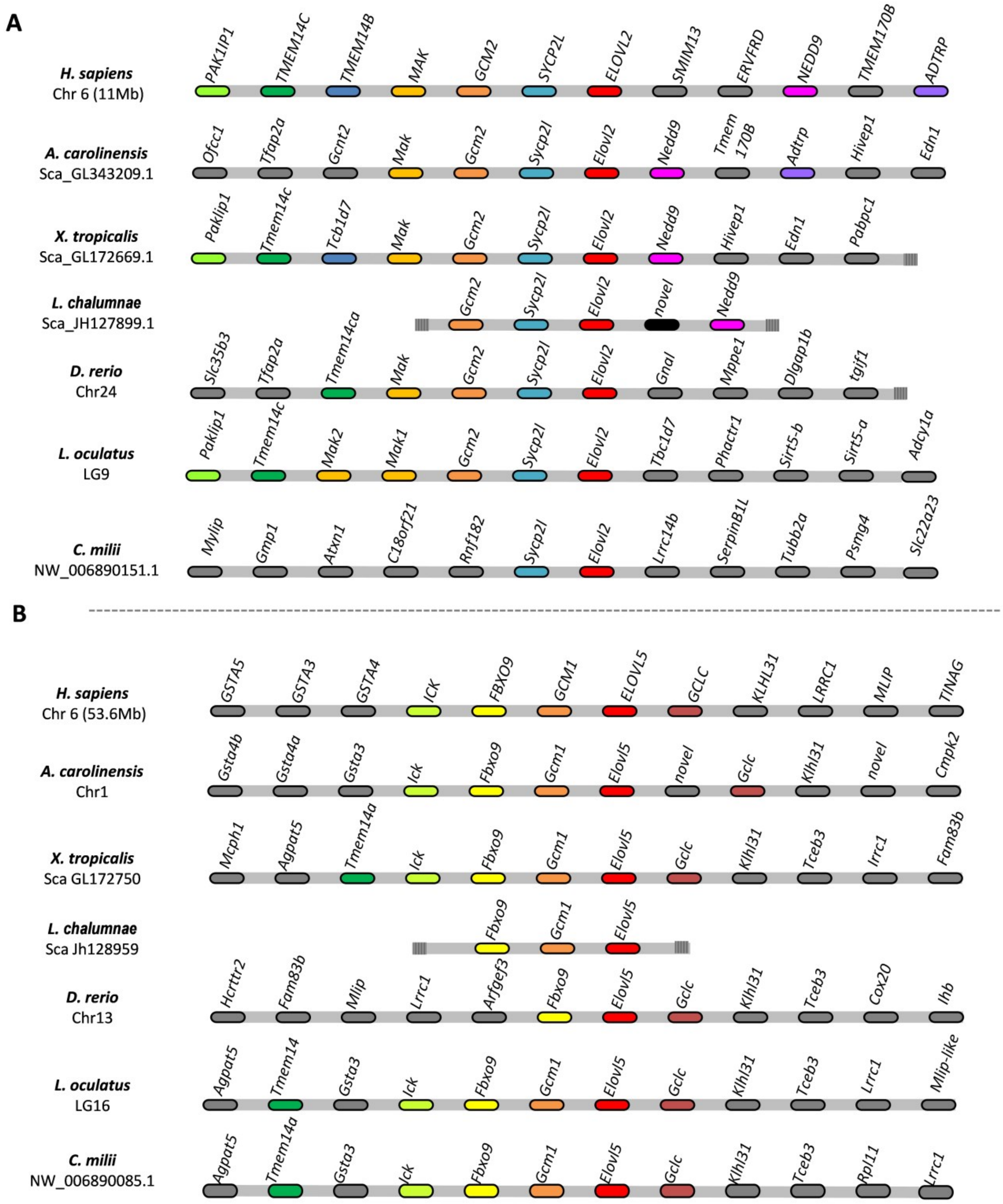


**Figure 3.** (A) Comparative synteny maps of the *Elovl2* locus and (B) *Elovl5* locus in *H. sapiens, A. carolinensis, X. tropicalis, L. chalumnae, D. rerio, L. oculatus* and *C. milii.*

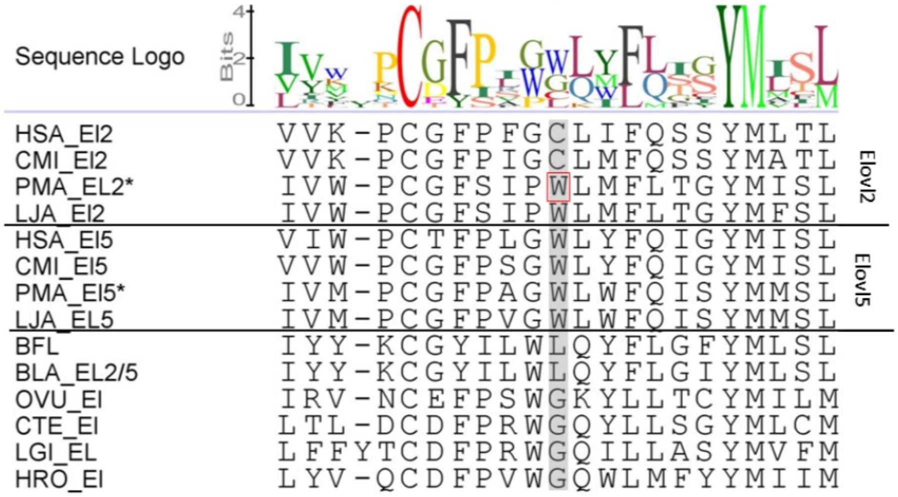

Sequence Logo
Bits
4
2
0
HSA_El2 VVK-PCGFPFGCLIFQSSYMLTL
CMI_El2 VVK-PCGFPIGCLMFQSSYMATL
PMA_EL2* IVW-PCGFSIPWLMFLTGYMISL
LJA_El2 IVW-PCGFSIPWLMFLTGYMFSL
HSA_El5 VIW-PCTFPLGWLYFQIGYMISL
CMI_El5 VVW-PCGFPSGWLYFQIGYMISL
PMA_El5* IVM-PCGFPAGWLWFQISYMMSL
LJA_EL5 IVM-PCGFPVGWLWFQISYMMSL
BFL IYY-KCGYILWLQYFLGFYMLSL
BLA_EL2/5 IYY-KCGYILWLQYFLGIYMLSL
OVU_El IRV-NCEFPSWGKYLLTCYMILM
CTE_El LTL-DCDFPRWGQYLLSGYMLCM
LGI_EL LFFYTCDFPRWGQILLASYMVFM
HRO_El LYV-QCDFPVWGQWLMFYYMIIM
Elovl2
Elovl5

**Figure 4.** Sequence alignment identifying the mutated site in lamprey. HSA – *H. sapiens;* CMI – *C. milii,* PMA - *P. marinus,* LJA – *L. japonicum;* BFL – *B. floridae;* BLA – *B. lanceolatum ;* OVU – *O. vulgaris;* CTE – *C. teleta;* LGI – *L. gigantea;* HRO - *H. robusta.* *- denotes sequences isolated and cloned in this work, red box highlight mutation site W>C.

## Supplementary Tables

**Table 1.** LC-PUFA in the biosynthetic cascade.

| ω3 | |
|---|---|
| **PUFA name** | **Symbol** |
| α-linolenic acid | 18:3n-3 |
| Stearidonic acid | 18:4n-3 |
| Eicosatetraenoic acid | 20:4n-3 |
| Eicosapentaenoic acid | 20:5n-3 |
| Docosapentaenoic acid | 22:5n-3 |
| Tetracosapentaenoic acid | 24:5n-3 |
| Tetracosahexaenoic acid | 24:6n-3 |
| Docosahexaenoic acid | 22:6n-3 |
| **ω6** | |
| **PUFA name** | **Symbol** |
| linoleic acid | 18:2n-6 |
| γ-linolenic acid | 18:3n-6 |
| Dihomo-γ-linolenic acid | 20:3n-6 |
| Arachidonic acid | 20:4n-6 |
| Adrenic acid | 22:4n-6 |
| Tetracosatetraenoic acid | 24:4n-6 |
| Tetracosapentaenoic acid | 24:5n-6 |
| Docosapentaenoic acid | 22:5n-6 |

**Table 2.** Accession numbers of all sequences used in phylogenetic analysis.

| Species | *Elovl2* | *Elovl5* | *Elovl4* |
|---|---|---|---|
| **Human** *(Homo sapiens)* | XP_011513019.1 | NP_068586.1 | NP_073563.1 |
| **Opossum** *(Monodelphis domestica)* | XP_007488013 | XP_001364339.1 | XP_001366145.1 |
| **Chicken** *(Gallus gallus)* | NP_001184237.1 | NP_001186126.1 | NP_001184238.1 |
| **Chinese softshell turtle** *(Pelodiscus sinensis)* | XP_006138890 | XP_006137802.1 | - |
| **Green Anole** *(Anolis carolinensis)* | NP_001016159.1 | XP_003215478.1 | XP_003215742.1 |
| **Western clawed frog** *(Xenopus tropicalis)* | NP_001016159.1 | NP_001011248.1 | - |
| **African coelacanth** *(Latimeria chalumnae)* | XP_006006450.1 | XP_006010670.1 | XP_006008610.1 |
| **Atlantic salmon** *(Salmo salar)* | NP_001130025 | NP_001117039.1 | - |
| **Rainbow trout (***Oncorhynchus mykiss)* | AIT56593.1 | - | - |
| **Northern pike** *(Esox lucius)* | XP_010884057.1 | - | - |
| **Zebrafish** *(Danio rerio)* | AAI29269.1 | NP_956747.1 | NP_957090.1<br>NP_956266.1 |
| **Medaka** *(Oryzias latipes)* | - | XP_004077464.1 | - |
| **Three-spined stickleback** *(Gasterosteus aculeatus)* | - | ENSGACT00000008538 | - |
| **Spotted gar** *(Lepisosteus oculatus)* | XP_006634635.1 | XP_006638754.1 | - |
| **European Eel** *(Anguilla anguilla)* | JAH99109 | - | - |
| **Elephant shark** *(Callorhynchus milii)* | XP_007900820/KT462565 | XP_007892243.1/KT462566 | - |
| **Small-spotted catshark** *(Scyliorhinus canicula)* | - | Transcript-ctg18611 | - |
| **Sea lamprey** *(Petromyzon marinus)* | KT462563 | KT462564 | S4R5D2 |
| **Japanese lamprey** *(Lethenteron japonicum)* | JL4990 | JL3695 | JL12276 |
| **Sea squirt** *(Ciona intestinalis)* | - | | NP_001029014.1 |
| **Florida amphioxus** *(Branchiostoma floridae)* | JGI_211218 | | - |
| **European amphioxus** *(Branchiostoma lanceolatum)* | KT462562 | | - |
| **Common octopus** *(Octopus vulgaris)* | AFM93779.1 | | - |
| **Polychaete worm** *(Capitella teleta)* | ELU18884.1 | | ELU05135.1 |
| **Owl limpet** *(Lottia gigantea)* | JGI protein Id 224291 | | - |
| **Freshwater leech** *(Helobdella robusta)* | JGI protein Id 63042 | | - |

**Supplementary table 3.** Details of all primers and PCR conditions.

| Specie | Gene | Primer F | Primer R | Tm | Function |
|---|---|---|---|---|---|
| *Branchiostoma lanceolatum* | *Elovl2/5* | TGGTACTACTTCTCCAAGGCCathgarttyyt | TGGGCCTGGGTGATGTACykyttccacca | 55 | Degenerate primers |
| | | CGCAGGATGAAGAACAACGTGTCA | GGCTAACTCGTTCATCCACGTCATC | 65 | GSP RACE primers |
| | | TGCACTACCCACCATACGAA | TTTCAAATCGGTCGGATAGG | 58 | ORF |
| | | CCCGGTACCAAGATGGCCACGACCACTGCAACTG | CCCTCTAGAGGTCATTCGGCTTTCTTAGCCCTCC | 65 | Cloning primers |
| *Petromyzon marinus* | *Elovl2* | - | CCGCCAGCCCGTAGTAGGAGTACAT | 60 | GSP RACE primers |
| | | GGTATCAACGCAGAGTACATGG | AGTTTTGGACTAATCGCGTCAC | 58 | ORF |
| | | CTTCTCCATCCCGTGTCTCATGTTCCTG | CAGGAACATGAGACACGGGATGGAGAAG | 62 | Site directed mutagenesis |
| | | CCCGGTACCACCATGGAATTCTTGGATAACACACTCAATG | CCCTCTAGAGAATCGCCTCAGTCCAGAGCAACC | 68 | Cloning primers |
| | *Elovl5* | GCGCTTATGCTTACTGAATGT | TGGCATTTCCTTCTCTTCCAAT | 58 | ORF |
| | | CCCGGATCCACAATGGAGGCACTGGACACAGC | CCCTCTAGATTACACGCGCTTGGGCTTGCGC | 68 | Cloning primers |

## IV.2 Molecular and functional characterization of a fads2 orthologue in the Amazonian teleost, *Arapaima gigas*

Mónica Lopes-Marques, Rodrigo Ozório, Ricardo Amaral, Douglas R. Tocher,

Óscar Monroig, L. Filipe C. Castro

**Comparative Biochemistry and Physiology, Part B**

(DOI: 10.1016/j.cbpb.2016.09.007)

# Molecular and functional characterization of a *fads2* orthologue in the Amazonian teleost, *Arapaima gigas*

Mónica Lopes-Marques [a,b], Rodrigo Ozório [a,b], Ricardo Amaral [c], Douglas R. Tocher [d], Óscar Monroig [d,*,1], L. Filipe C. Castro [a,e,*,1]

[a] CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, U. Porto – University of Porto, Porto, Portugal
[b] ICBAS - Institute of Biomedical Sciences Abel Salazar, U. Porto - University of Porto, Portugal
[c] Universidade Federal do Acre, Brazil
[d] Institute of Aquaculture, Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK
[e] Department of Biology, Faculty of Sciences, U. Porto - University of Porto, Portugal

ARTICLE INFO

*Article history:*
Received 8 June 2016
Received in revised form 8 September 2016
Accepted 27 September 2016
Available online 29 September 2016

*Keywords:*
*Arapaima gigas*
Fatty acid desaturase (Fads)
Long-chain polyunsaturated fatty acids (LC-PUFAs)
Teleosts
Evolution

ABSTRACT

The Brazilian teleost *Arapaima gigas* is an iconic species of the Amazon. In recent years a significant effort has been put into the farming of arapaima to mitigate overfishing threats. However, little is known regarding the nutritional requirements of *A. gigas* in particular those for essential fatty acids including the long-chain polyunsaturated fatty acids (LC-PUFA) eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA). The ability to biosynthesize LC-PUFA is dependent upon the gene repertoire of fatty acyl desaturases (Fads) and elongases (Elovl), as well as their fatty acid specificities. In the present study we characterized both molecularly and functionally an orthologue of the desaturase fatty acid desaturase 2 (*fads2*) from *A. gigas*. The isolated sequence displayed the typical desaturase features, a cytochrome $b_5$-domain with the heme-binding motif, two transmembrane domains and three histidine-rich regions. Functional characterization of *A. gigas fads2* showed that, similar to other teleosts, the *A. gigas fads2* exhibited a predominant Δ6 activity complemented with some capacity for Δ8 desaturation. Given that *A. gigas* belongs to one of the oldest teleostei lineages, the Osteoglossomorpha, these findings offer a significant insight into the evolution LC-PUFA biosynthesis in teleosts.

## 1. Introduction

Long-chain polyunsaturated fatty acids (LC-PUFA) play vital roles in numerous biological processes. They participate in structural functions as major components of biomembranes and are also involved in processes such as the inflammatory response, reproduction (Wall et al., 2010; Robinson and Mazurak, 2013), neural development (Perica and Delaš, 2011) and can have beneficial effects in pathological conditions such as cardiovascular disease (Psota et al., 2006; Jump et al., 2012). LC-PUFA are often defined as compounds with 20 to 24 carbon atoms and three or more double bonds (unsaturations), and can be classified into two main groups: the omega-6 (ω6 or n-6) and the omega-3 (ω3 or n-3) LC-PUFA, based upon the position of the first double bond in relation to the methyl end carbon ($CH_3$) (Monroig et al., 2011a). LC-PUFA of the n-6 and n-3 series can be of dietary origin or, alternatively, they can be biosynthesized from dietary essential fatty acids (EFA) such as linoleic acid (LA, 18:2n-6) and α-linolenic acid (ALA, 18:3n-3), respectively, through a series of sequential biochemical reactions, mediated by elongation of very long-chain fatty acid protein (Elovl) and fatty acyl desaturases (Fads).

The ability to endogenously synthesize LC-PUFA from dietary fatty acids (FA) differs markedly among vertebrate species (Rivers et al., 1975; Bauer, 1997; Tocher, 2003; Burdge and Calder, 2005; Fonseca-Madrigal et al., 2014; Castro et al., 2016; Monroig et al., 2016a, 2016b). This variation may be primarily attributed to differences in the *elovl* and *fads* gene repertoire, as well as their associated fatty acid substrate specificities. For instance, mammals have several *FADS* genes of which *FADS1* encodes a Δ5 desaturase and *FADS2* encodes a desaturase with Δ6 preference, in addition to Δ4 activity reported in some mammals (Park et al., 2009, 2015). In contrast, teleost fish examined to date have been found to possess exclusively *FADS2* orthologues (Castro et al., 2012, 2016). However, while mammalian FADS enzymes are essentially mono-functional, mechanisms of bifunctionalization (i.e., acquisition of additional/alternative substrate specificities) have been described in several teleost Fads2. Thus, Fads2 with dual Δ6Δ5 desaturase activities have been described in *Danio rerio* (Hastings et al., 2001), *Siganus canaliculatus* (Li et al., 2010), *Oreochromis niloticus* (Tanomman et al., 2013), *Chirostoma estor* (Fonseca-Madrigal et al., 2014) and *Clarias gariepinus* (Oboh et al., 2016). In addition,

[*] Corresponding author.
*E-mail addresses:* oscar.monroig@stir.ac.uk (Ó. Monroig), filipe.castro@ciimar.up.pt (L.F.C. Castro).
[1] Contributed equally to this work.

http://dx.doi.org/10.1016/j.cbpb.2016.09.007
1096-4959/© 2016 Elsevier Inc. All rights reserved.

*S. canaliculatus* and *C. estor* possess a duplicated *Fads2* that exhibit Δ4 desaturase activity (Li et al., 2010; Fonseca-Madrigal et al., 2014), a type of enzyme also found in *Solea senegalensis* (Morais et al., 2012) and *Channa striata* (Kuah et al., 2015). Moreover, in agreement with the abilities reported in the baboon Δ6-desaturase (Park et al., 2009), the majority of teleost Fads2 desaturases have been demonstrated to possess the capability for Δ8 desaturation (Monroig et al., 2011b). Overall the complement of LC-PUFA biosynthetic enzymes, namely FADS and ELOVL, as well as their functionalities, dictates the ability of a species for the conversion of $C_{18}$ PUFA (LA and ALA) into physiologically important LC-PUFA including arachidonic acid (ARA, 20:4n-6), eicosapentaenoic acid (EPA, 20:5n-3) and docosahexaenoic acid (DHA, 22:6n-3) (Bell and Tocher, 2009; Castro et al., 2016). Importantly, the investigation of Fads and Elovl in fish has primarily focused on farmed species since both Fads and Elovl capabilities underpin the efficiency of these fish species to utilize the $C_{18}$ PUFA present in vegetable oils (VO) currently used as sustainable replacements for dietary fish oils (FO) in aquafeeds (Tocher, 2010). Therefore a clear understanding of LC-PUFA biosynthesis pathways is critical to understand the potential limitations of farmed fish species and for the implementation of dietary strategies to fulfil essential requirements and ensure normal growth and development in captivity.

An iconic species of the Amazon, so-called "pirarucú" (*Arapaima gigas*), is one of the largest freshwater and air-breathing fishes in the world, and has been extensively fished since the 18th century (Veríssimo, 1895; Goulding, 1980). In the early 1970's over-exploitation of *A. gigas* led to its near extinction (Goulding, 1980) and listing in CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora). To overcome this threat, considerable effort has been put into developing the sustainable farming of this species. However, despite some important advances, critical knowledge in key areas such as physiology and nutrition is still scarce in this species. Much of the published research on *A. gigas* has focused on the understanding and evolution of the air-breathing capacity (Brauner et al., 2004; Gonzalez et al., 2010), general health and aquaculture practices (Ribeiro et al., 2011; Bezerra et al., 2014) and, more recently, the potential use of *A. gigas* scales as biomaterials (Torres et al., 2015). In contrast, few studies have addressed the dietary requirements of *A. gigas* (Ituassú et al., 2005; Andrade et al., 2007; Ribeiro et al., 2011), stressing the need for a broader understanding of the metabolism of this carnivorous species. Here, we describe the isolation and functional characterization of a cDNA from *A. gigas* orthologous to *fads2* desaturases, key enzymes in LC-PUFA biosynthetic pathways and crucial elements in determining EFA requirements in this species. The phylogenetic position of *A. gigas* within one of the most ancient teleost lineages, the Osteoglossomorpha, brings new insights into the evolution of the LC-PUFA biosynthesis cascade in both fish and vertebrates in general.

## 2. Materials and methods

### 2.1. Molecular cloning of the A. gigas fads gene

Total RNA was extracted from a range of *A. gigas* tissues using the Illustra RNAspin Mini kit (GE Healthcare, UK). The RNA extraction process included an on-column DNase I treatment (provided in the kit). RNA integrity was assessed on a 1% agarose TAE gel stained with GelRed™ nucleic acid stain (Biotium, Hayward, CA, USA). The Quant-iT™ RiboGreen® RNA Assay Kit (Life Technologies, Carlsbad, CA, USA) was used to measure total RNA concentration. Reverse transcription reactions were performed with the iScript cDNA Synthesis Kit (Bio-Rad, Hercules, CA, USA).

*Arapaima gigas* FADS gene was isolated in three main steps. First, degenerate primers targeting the Fads gene were designed using CODEHOP (Rose et al., 2003) available at http://blocks.fhcrc.org/codehop.html. The initial polymerase chain reaction (PCR) was performed with a degenerate primer set and Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific, Waltham, USA), set for a final volume of 20 µl, with 500 nM of sense and antisense primers, and 1 µl of *A. gigas* cDNA pool (see Table 1 for primers, PCR conditions). In the second step, the partial *fads* sequence was further extended by Rapid amplification of cDNA ends (RACE) PCR using as template 5′ and 3′ RACE ready cDNA prepared with SMARTer™ RACE cDNA Amplification Kit (Clontech, CA, USA). Gene specific primers for RACE were designed using the previously isolated fragment and RACE PCR was performed with Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific) using 1 µl of gene specific primer combined with 2 µl Universal primer mix (Clontech) (see Table 1 for primers and PCR conditions). The resulting 5′and 3′sequences were assembled to produce the full open reading frame (ORF) *fads*-like cDNA. In the final step, the full ORF of *A. gigas* FADS was isolated using 1 µl of *A. gigas* cDNA pool, and Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific, Waltham, USA), set for a final volume of 20 µl, with 500 nM of sense and antisense primers (see Table 1 for primers and PCR conditions). In each step resulting PCR products were analyzed in 1% agarose gel, purified with NZYGelpure (NZYTech, Lisbon, Portugal) and confirmed by sequencing (GATC Biotech Constance, Germany). The final, full ORF sequence was translated and submitted to pFAM and NCBI for blastp searches retrieving Fads-like profile (Accession number: KX809739).

### 2.2. Sequence collection, phylogenetic and 2D structural analysis

Fads amino acid (aa) sequences were retrieved from Genbank and Ensembl (for accession numbers see Table 2). Sequences were aligned with MAFFT using the L-INS-i method (Katoh and Toh, 2008). The sequence alignment was stripped from all columns containing gaps leaving 374 gap-free sites for phylogenetic analysis. Maximum likelihood phylogenetic analysis was performed in PhyML v3.0 server (Guindon et al., 2010) using smart model selection resulting in LG + G + I + F, and branch support was calculated using 1000 bootstraps. Using the same alignment a second Bayesian phylogenetic analysis was performed using MrBayes v3.2.3 available in CIPRES Science Gateway V3.3 (Miller et al., 2015). MrBayes was run for 1 million generations with the following parameters: rate matrix for aa = mixed, nruns = 2, nchains = 4, temp = 0.2, sampling set to 500 and burin to 0.25. The resulting trees were visualized in Fig Tree V1.3.1 available at http://tree.bio.ed.ac.uk/software/figtree/ and rooted at mid-point. *A. gigas* aa sequence was submitted to TOPCONS web server for prediction of 2D topology, with all parameters set to default (http://topcons.net/) (Tsirigos et al., 2015), and results visualized using Potter web application (http://wlab.ethz.ch/protter) (Omasits et al., 2014).

### 2.3. Yeast expression assays and fatty acid analysis

The *A. gigas fads* ORF was isolated with two sequential PCR with Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific, USA) as described above. The first PCR was performed with an *A. gigas* cDNA pool and primers (AgigasFADS_ORF_F and AgigasFADS_ORF_R, Table 1) targeting the full ORF. The PCR product was diluted (1:50) and used as template for the second PCR performed with primers containing restriction sites for *Kpn*I (AgigasFADS_pYES_KpnI_F) and *Xba*I (AgigasFADS_pYES_XbaI_R) (Table 1). The final PCR product was purified and digested with the appropriate restriction enzymes and cloned into the yeast expression vector pYES2 (Invitrogen, CA, USA). Transformation and culture of yeast *Saccharomyces cerevisiae* were conducted as previously described (Hastings et al., 2001; Agaba et al., 2004; Oboh et al., 2016). Briefly, transgenic yeast expressing the *A. gigas fads* ORF were grown in the presence of PUFA including Δ6 (18:3n-3 and 18:2n-6), Δ8 (20:2n-6 and 20:3n-3), Δ5 (20:4n-3; 20:3n-6) and Δ4 (22:5n-3 and 22:4n-6) desaturase substrates. PUFA substrates, added as sodium salts, were supplemented in the yeast medium at final concentrations of 0.5 mM ($C_{18}$), 0.75 mM ($C_{20}$) and 1.0 mM ($C_{22}$) as uptake efficiency decreases with increasing chain

**Table 1**
Primer sets and corresponding PCR conditions.

| Primer set function | Primer name | Primer sequence | Initial denaturation | Cycles | Denaturation | TM | Extension (size bp) | Final extension |
|---|---|---|---|---|---|---|---|---|
| Degenerate primers | FADS2degen_F<br>FADS2degen_R | GCGCCTCCGCCAAytggtggaayc<br>TGGCCGGAGAACcartcrttraa | 98 °C/10 s | 40 | 98 °C/1 s | 54 °C/5 s | 72 °C/10 s | 72 °C/1 min |
| Gene specific race primers | 3RC_AgigasFADS_F<br>5RC_AgigasFADS_R | ACCTAAAGGGTGCTTCAGCCAACT<br>GTTCGGAACAAGCCCTCTTTCTC | 98 °C/10 s | 20 | 98 °C/1 s | 62 °C/5 s | 72 °C/15 s | 72 °C/1 min |
| Nested gene specific race primers | N3RC_AgigasFADS_F<br>N5RC_AgigasFADS_R | GTTTCTGGAGAGCCACTGGTTTGT<br>CTGCGTTTTTCTGGCGGTCTAAG | 98 °C/10 s | 35 | 98 °C/1 s | 62 °C/5 s | 72 °C/8 s | 72 °C/1 min |
| Full ORF | AgigasFADS_ORF_F<br>AgigasFADS_ORF_R | ATATTGCCAGAGGATGGATG<br>GGGCCTCATTACATTCAATAAA | 98 °C/10 s | 20 | 98 °C/1 s | 56 °C/5 s | 72 °C/22 s | 72 °C/1 min |
| Restriction site primers for cloning | AgigasFADS_pYES_KpnI_F<br>AgigasFADS_pYES_XbaI_R | CCCGGTACCAAGATGGCGGCGGGGGGCA<br>CCCTCTAGAGGGGTTACTTGTGGAGATA CGCATC | 98 °C/10 s | 35 | 98 °C/1 s | 67 °C/5 s | 72 °C/20 s | 72 °C/1 min |

length (Zheng et al., 2009). After 48 h of incubation, yeast were harvested, washed and total lipid extracted by homogenization in chloroform/methanol (2:1, v/v) containing 0.01% BHT (Monroig et al., 2013). Fatty acyl methyl esters (FAME) were prepared from total lipids extracted from harvested cells and identified based on GC retention times and confirmed by GC–MS as described previously (Hastings et al., 2001; Li et al., 2010). FA desaturation efficiencies from exogenously added PUFA substrates were calculated by the proportion of substrate FA converted to a desaturated product as (product area/(product area + substrate area)) × 100.

## 3. Results

### 3.1. Sequence conservation and topology prediction

The isolated *A. gigas* sequence was translated and submitted to BLASTp and to PFam to validate the *fads*-like profile and identify the main protein domains. BLASTp searches showed that the *A. gigas* sequence had highest identity scores with *fads2* desaturases from other teleost species (results not shown), while the PFam search identified two main domains typical of Fads enzymes: a cytochrome $b_5$-like heme/steroid binding domain (15 - 88 aa) and FA desaturase domain (150 - 412 aa). To further characterize, the *A. gigas* Fads-like protein was aligned with four known and fully characterized Fads aa sequences from *D. rerio* (NCBI Protein accession no Q9DEX7.1), *Salmo salar* (NCBI Protein accession no NP_001117047.1), *O. niloticus* (NCBI Protein accession no AGV52807.1) and *Homo sapiens* (NCBI Protein accession no NP_004256.1) (Fig. 1A). The *A. gigas* sequence showed highest degree of pairwise identity with the *S. salar* Fads2 (86.1%), followed by Fads2 from *O. niloticus* (83.9%), *D. rerio* (82.8%) and *H. sapiens* (79.3%), revealing a high degree of cross-species conservation. Additionally, using *H. sapiens* FADS2 sequence as a reference, several sequence signature motifs of Fads enzymes were identified: the heme binding motif HPGG and three histidine boxes HXXXH, HXXHH and QXXHH, which are presumed to form the Fe-binding active center of the enzyme (Los and Murata, 1998; Pereira et al., 2003) (Fig. 1A). The heme binding motif was totally conserved in Fads from all species analyzed including *A. gigas*. In the first histidine box two distinct patterns were observed: HDYGH in *H. sapiens* and *S. salar*, while *A. gigas*, *D. rerio* and *O. niloticus* showed the signature HDFGH with the replacement of a tyrosine (Y) by a phenylalanine (F) (Fig. 1A). In the second histidine box, all analyzed species presented HFQHH with the exception of *O. niloticus*, whose Fads2 presents HFRHH (Fig. 1A). Full conservation of the third histidine box was found across all the analyzed species.

**Table 2**
Accession number of sequences used phylogenetic analysis.

| Species | Accession number<br>FADS2 | FADS1 |
|---|---|---|
| HSA- *Homo sapiens* | NP_004256.1 | NP_037534.3 |
| MDO- *Monodelphis domestica* | – | H9H609 |
| ASI- *Alligator sinensis* | XP_006033391.1 | XP_006033402.1 |
| GGA- *Gallus gallus* | NP_001153900.1 | XP_421052.4 |
| LCH- *Latimeria chalumnae* | XP_005988034.1 | XP_005988035.1 |
| CMI - *Callorhinchus milii* | XP_007885636.1 | XP_007885635.1 |
| SCA- *Scyliorhinus canicula* | AEY94455.1 | – |
| DRE- *Danio rerio* | NP_571720.2 | – |
| SSA- *Salmo salar* | NP_001117047.1 | – |
| ONI-*Oreochromis niloticus* (a) | XP_005470661.1 | – |
| ONI-*Oreochromis niloticus* (b) | XP_003440520.1 | – |
| TMA - *Thunnus maccoyii* | ADG62353.1 | – |
| GMO - *Gadus morhua* | AAY46796 | – |
| BFL - *Branchiostoma floridae* | XP_002586930.1 | |

Regarding the 2D topology prediction, all calculation methods were consistent in predicting that *A. gigas* Fads-like displayed four membrane spanning domains, and that the N- and the C-terminals, as well as the three histidine motifs, were oriented towards the cytosol (Supplementary Material 1). Interestingly, the residues involved in regioselectivity were localized at the base of the third membrane spanning domain (Fig. 1B). The topology predicted for the *A. gigas* Fads2 was thus consistent with the structural organization proposed in previous reports for other Fads-like desaturases (Los and Murata, 1998; Meesapyodsuk et al., 2007; Lim et al., 2014).

### 3.2. Phylogenetic analysis of Fads-like ORF from A. gigas

Two phylogenetic analyses were conducted using the same data set consisting of aa sequence alignment between the newly cloned *A. gigas* putative Fads with FADS1 and FADS2 desaturase sequences from eighteen vertebrate species (mammals - *H. sapiens*, *M. domestica* birds – *G. gallus*, reptiles - *A. sinensis*, coelacanth - *L. chalumnae*, teleosts - *G. morhua*, *T. maccoyii*, *O. niloticus*, *S. salar*, and *D. rerio*, chondrichthyans - *S. canicula*, *C. milii* and one invertebrate (*B. floridae*). In both cases the tree topology showed two well-supported clades, one corresponding to the FADS1 and the second corresponding to the FADS2, being both trees out grouped by invertebrate FADS from *B. floridae*. The *A. gigas* Fads-like sequence strongly grouped (930 bootstraps or 1 posterior probabilities) (See Fig. 2) within the teleost group composed of all Fads2 sequences. Out grouping the teleost clade we find tetrapod and chondrichthyans Fads2 desaturases, indicating that the *A. gigas* putative Fads is a true *fads2* orthologue (See Fig. 2). However, desaturases with different substrate preferences, for example *D. rerio* and *O. niloticus* Fads2 that are bifunctional Δ6Δ5 desaturases (Hastings et al., 2001; Tanomman et al., 2013), and *G. morhua* and *S. salar* Fads2 that have been reported as unifunctional Δ6 desaturases (Zheng et al., 2005; Monroig et al., 2010) were found within the teleost clade.

### 3.3. Functional analysis of Fads2 in A. gigas

Functional characterization of the *A. gigas* desaturase was performed with using a well-established heterologous system consisting of yeast *S. cerevisiae* expressing the ORF of the *A. gigas fads2* and grown in the

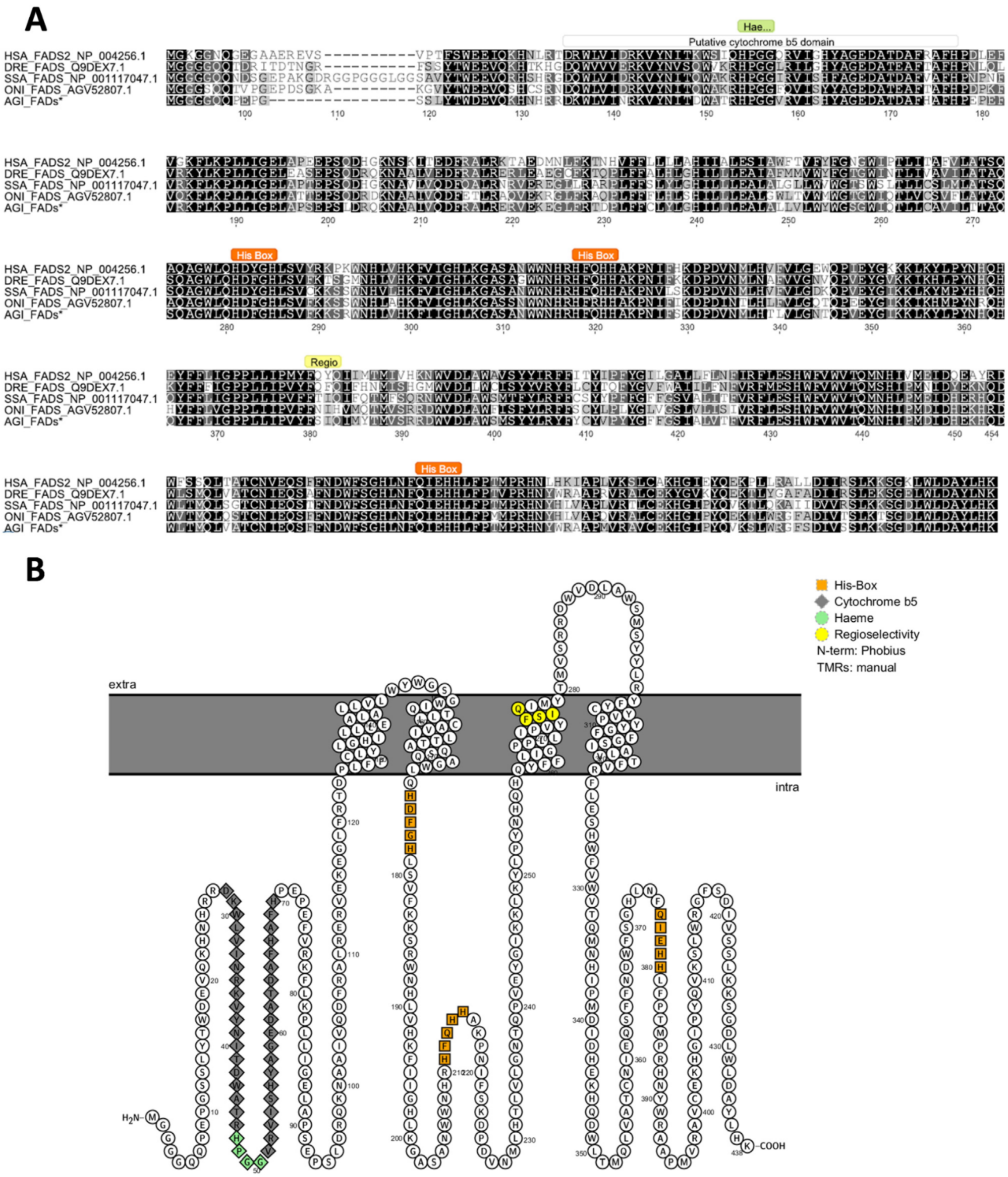


**Fig. 1.** Sequence analysis of *Arapaima gigas* Fads2. A, FADS sequence alignment, white: Cytochrome b5-like domain, green: heme binding motif, orange: conserved histidine boxes, and yellow reported regioselectivity residues. B, Predicted 2D topology of *Arapaima gigas* Fads color code is maintained. Intra-Cytosol and Extra-Lumen. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

presence of potential desaturase PUFA substrates (Hastings et al., 2001; Agaba et al., 2004; Fonseca-Madrigal et al., 2014). FA profile of yeast transformed with the empty pYES2 plasmid (control) consisted of the yeast endogenous FA including 16:0, 16:1 isomers (16:1n-9 and 16:1n-7), 18:0, and 18:1 isomers (18:1n-9 and 18:1 n-7) and whichever exogenously PUFA substrate was added (data not shown). These results confirmed that the yeast endogenous enzymes were not active on the exogenously added PUFA substrates (Agaba et al., 2005). On the other hand, yeast transformed with the ORF of the *A. gigas fads2* showed additional peaks when grown in the presence of 18:3n-3, 18:2n-6, 20:3n-3 and 20:2n-6 (Fig. 3). Thus, transgenic yeast expressing the *fads2* had the ability to desaturate 18:3n-3 and 18:2n-6 to 18:4n-3

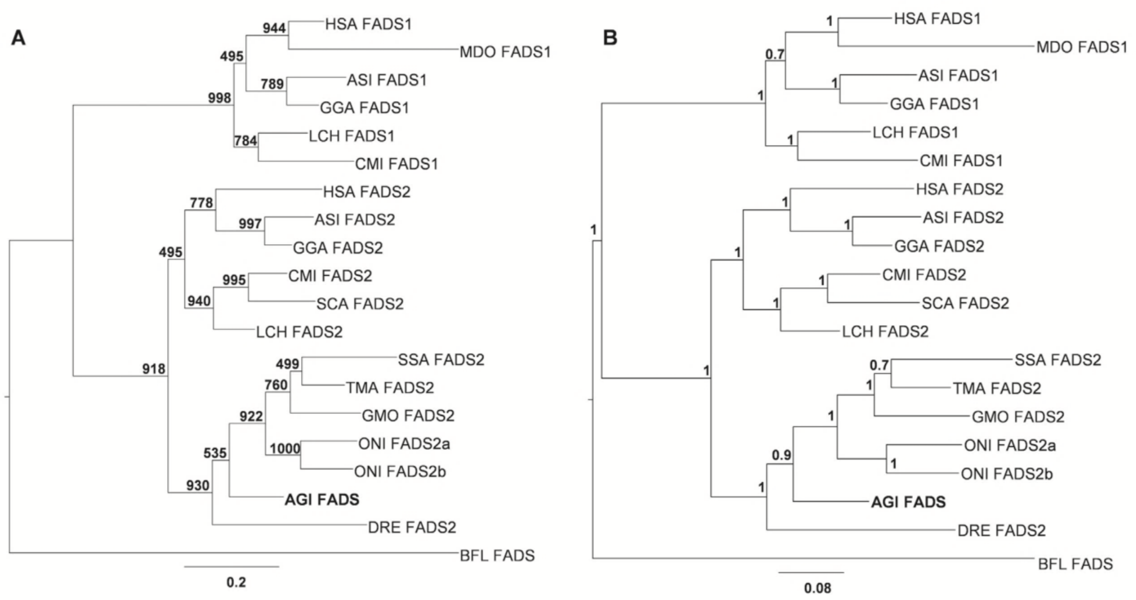


**Fig. 2.** Molecular phylogenetic analysis. A - Maximum likelihood phylogenetic analysis, node values indicate bootstrap replicates; B – Bayesian phylogenetic analysis node values indicate posterior probabilities. HSA- *Homo sapiens*, MDO - *Monodelphis domestica*, GGA - *Gallus gallus*, ASI - *Alligator sinensis*, LCH - *Latimeria chalumnae* DRE- *Danio rerio*; AGI- *Arapaima gigas*; ONI - *Oreochromis niloticus*; SSA- *Salmo salar*; GMO - *Gadus morhua*; TMA- *Thunnus maccoyii*; CMI - *Callorhinchus milii*, SCA- *Scyliorhinus canícula*. BFL – *B. floridae*.

(Fig. 3A) and 18:3n-6 (Fig. 3B), respectively, showing this enzyme has Δ6 desaturase activity. Moreover, transgenic yeast supplemented with 20:3n-3 and 20:2n-6 produced additional peaks identified as 20:4n-3 (Fig. 3C) and 20:3n-6 (Fig. 3D), respectively, showing that the *A. gigas fads2* had also Δ8 desaturase activity. Therefore, the data confirmed that the cloned *A. gigas fads2* encoded an enzyme with Δ6 and Δ8 desaturase specificities. Conversions obtained in the yeast expression system suggested that the *A. gigas* Fads2 has Δ6 as the most prominent activity and a preference for n-3 fatty acid substrates compared with n-6 substrates for each homologous FA substrate pair (Δ6 or Δ8) considered (Table 3). Neither Δ5 nor Δ4 activities were detected in yeast (Fig. 3E–H).

## 4. Discussion

Fads are, together with Elovl, key enzymes in LC-PUFA biosynthetic pathways (Castro et al., 2016; Monroig et al., 2016b)). The sequential and concerted action of both enzymes defines the ability of a given species to endogenously synthesize physiologically relevant LC-PUFA including ARA, EPA or DHA (Bell and Tocher, 2009). The investigation of the molecular components of LC-PUFA biosynthetic pathway in fish has been an active field of research over the last decade (Agaba et al., 2005; Zheng et al., 2009; Monroig et al., 2011b, 2012; Castro et al., 2012, 2016; Carmona-Antonanzas et al., 2013). This is particularly true in farmed fish species where a full understanding of LC–PUFA biosynthesis capacities is crucial to successfully grow fish on diets that are necessarily being formulated with ever-increasing levels of VO (rich in $C_{18}$ PUFA but devoid of LC–PUFA) as primary lipid sources to replace FO (Turchini et al., 2009). Overall, these studies have highlighted a surprisingly diverse and interesting pattern among Fads substrate specificities (Fonseca-Madrigal et al., 2014).

The primary objective of the present study was the molecular cloning and functional characterization of a desaturase of the Amazonian teleost *A. gigas*. This freshwater species with aquaculture potential (Cavero et al., 2003) has been barely investigated in terms of nutritional requirements. In addition, *A. gigas* belongs to the Osteoglossiformes, a teleost order that has been considered to be the most basal of living teleosts (Nelson, 1994), therefore bringing a fresh perspective on the functional diversification of the desaturases in teleosts. The isolated Fads2 sequence of *A. gigas* showed all the typical features of fatty acyl (also known as "front-end") desaturases when subjected to BLASTp and to PFam searches. Furthermore, detailed sequence alignment analysis revealed that the unique structure of Fads-like enzymes was preserved in *A. gigas* Fads2 that contained three highly conserved histidine boxes, as well as the heme motif within the cytochrome $b_5$-like domain, which are considered to be involved in the formation of the desaturase catalytic centre (Shanklin et al., 1994; Los and Murata, 1998; Tocher et al., 1998). The 2D topology analysis of *A. gigas* Fads2 predicted four transmembrane domains TM1: 124-145, TM2: 151-172, TM3: 258-279, TM4: 300-321, that oriented the three histidine boxes and the cytochrome $b_5$-like domain to the cytosol, consistent with the structural organization proposed in previous reports (Los and Murata, 1998; Meesapyodsuk et al., 2007; Lim et al., 2014). Among the three histidine boxes, two distinct patterns were observed in the first histidine box in the Fads2, with *A. gigas*, *D. rerio* and *O. niloticus* having the signature HD$\underline{\text{F}}$GH, whereas a replacement of a phenylalanine (F) by tyrosine (Y) occurs for *H. sapiens* and *S. salar* Fads2. This replacement was predicted to not affect the mandatory/canonical histidine residues within each box. Additionally the abovementioned aa substitution was not expected to have any major functional impact, possibly due to the fact that these two aa residues share very similar biochemical properties (Betts and Russell, 2003). In contrast, differences were found within the residues previously proposed to participate in the regioselectivity of these enzymes (Hsa: 279Phe - 282Gln; Dre: 279Phe - 282Gln, Oni: 280Phe - His283, Ssa: 289Phe-292Gln; Agi: 273Phe - 276Gln) (Meesapyodsuk et al., 2007; Lim et al., 2014), possibly accounting for the different Fads activities observed in these species.

All *fads* characterized so far from teleosts are orthologous to *FADS2*, which performs primarily Δ6 desaturations in mammals (Guillou et al., 2010). This is further supported by the herein phylogenetic analysis of *A. gigas fads*, together with phylogenetic analyses reported previously (Zheng et al., 2004; Monroig et al., 2011b; Liu et al., 2014). However, the teleost Fads exhibit a wide range of PUFA specificities (Hastings et al., 2001, 2004; Li et al., 2010; Monroig et al., 2012; Xie et al., 2014), underscoring a "*functional plasticity*" that has been previously attributed as a consequence of adaptation to availability of

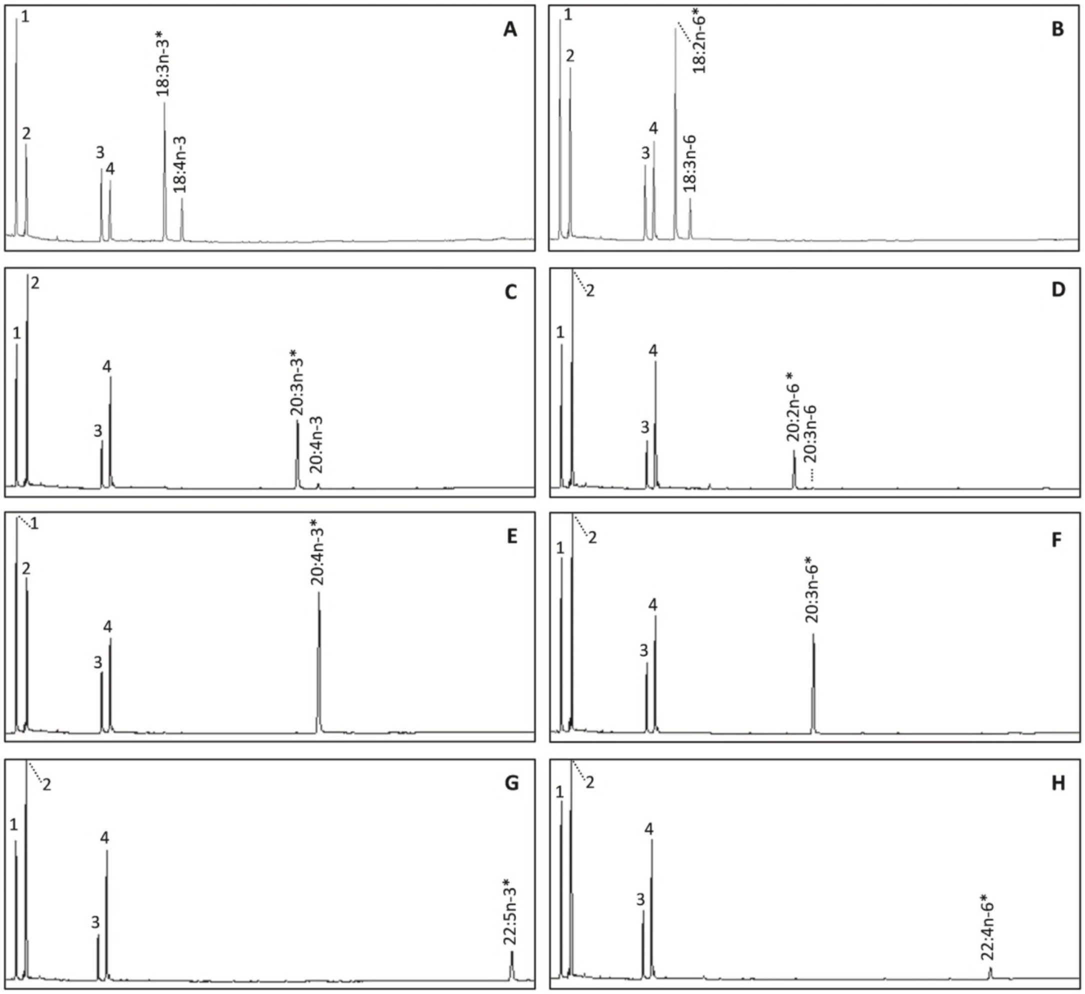


**Fig. 3.** Functional characterization of *Arapaima gigas* Fads2 in yeast (*Saccharomyces cerevisiae*). Fatty acid (FA) profiles were determined after the yeast were grown in the presence of exogenously added substrates indicates in each case by (*). Peaks 1–4 in all panels correspond to yeast endogenous FA, namely 1 - (16:0), 2 - (16:1n-7), 3 – (18:0) and 4 – (18:1n-9). FA derived from the exogenously added substrates or elongation products are indicated accordingly in each panel above the corresponding product.

LC-PUFA in variable habitats and trophic levels (Tocher, 2010; Monroig et al., 2011b, 2012; Castro et al., 2012; Fonseca-Madrigal et al., 2014). Thus, Fads2 with dual Δ6Δ5 activity have been cloned from *D. rerio* (Hastings et al., 2001), *S. canaliculatus* (Li et al., 2010), *O. niloticus* (Tanomman et al., 2013), *C. estor* (Fonseca-Madrigal et al., 2014), and *C. gariepinus* (Oboh et al., 2016). Moreover, teloest Fads2 with Δ4 desaturase activity have been found in *S. canaliculatus* (Li et al., 2010), *S. senegalensis* (Morais et al., 2012) and *C. striata* (Kuah et al., 2015). Interestingly, the human *FADS2* gene product has been recently demonstrated to have the ability for direct Δ4 desaturation of 22:5n-3 to 22:6n-3 (Park et al., 2015). Nevertheless, the majority of functionally characterized teleost Fads2 are essentially Δ6 desaturase enzymes as reported in a variety of teleost fish species including gilthead seabream, rainbow trout, Atlantic salmon (three genes), turbot, cobia, European seabass, barramundi, black seabream, nibe croaker, Northern bluefin tuna, meagre, Japanese eel and orange spotted grouper (Castro et al., 2016). In agreement, the *A. gigas* Fads2 was demonstrated to be a Δ6 desaturase able to convert 18:3n-3 and 18:2n-6 to 18:4n-3 and 18:3n-6, respectively.

**Table 3**
Functional characterization of the *Arapaima gigas* Fads2 in yeast. Conversions were calculated according to the formula (product area/(product area + substrate area)) × 100.

| FA substrate | FA product | % conversion |
|---|---|---|
| 18:3n-3 | 18:4n-3 | 25.8 |
| 18:2n-6 | 18:3n-6 | 16.1 |
| 20:3n-3 | 20:4n-3 | 5.8 |
| 20:2n-6 | 20:3n-6 | 3.8 |
| 20:4n-3 | 20:5n-3 | nd |
| 20:3n-6 | 20:4n-6 | nd |
| 22:5n-3 | 22:6n-3 | nd |
| 22:4n-6 | 22:5n-6 | nd |

nd, not detected.

However, in addition, the *A. gigas* Fads2 showed capability for Δ8 desaturation, since it was capable of converting both 20:3n-3 and 20:2n-6 into 20:4n-3 and 20:3n-6, respectively. This activity was first reported in the baboon FADS2 (Park et al., 2009) and subsequently described in a range of fish Fads2 enzymes (Monroig et al., 2011b). The capability for Δ8 desaturation appears widespread in Fads2 characterized from fish (Monroig et al., 2011b, 2013; Wang et al., 2014; Kabeya et al., 2015; Oboh et al., 2016), with few exceptions represented by the Atlantic salmon and rainbow trout Δ5 Fads2, as well as the striped snakehead

Δ4 Fads2 (Monroig et al., 2011b; Kuah et al., 2015; Abdul Hamid et al., 2016). Interestingly, it appeared that, generally, Fads2 from marine teleosts had relatively high Δ8 desaturase ability compared to their freshwater and salmonid counterparts (Monroig et al., 2011b). Consequently, the Δ6:Δ8 desaturation ratio varies among teleost Fads2, with marine species having relatively low Δ6:Δ8 ratios, while freshwater and salmonid species having higher Δ6:Δ8 ratios. The *A. gigas* Fads2 had a Δ6:Δ8 ratio of 4.4 for n-3 PUFA substrates (25.8:5.8), and thus more within the range of marine teleosts such as turbot (4.2) or gilthead seabream (2.7) and far from freshwater species like rainbow trout (91.5) and zebrafish (22.4). While it is unclear what the evolutionary drivers are for the high capacity for Δ8 desaturation in *A. gigas* Fads2, having a Fads2 with the ability to operate as a Δ6 desaturase on ALA and LA, and as a Δ8 on 20:3n-3 and 20:2n-6, may confer an advantage to this species enabling production of 20:4n-3 and 20:3n-6, respectively, through two different pathways. Both 20:4n-3 and 20:3n-6 are substrates of Δ5 desaturase, an enzyme that, despite being absent in the vast majority of teleosts, is likely to be retained in basal teleosts such as Osteoglossidae, the family to which *A. gigas* belongs. In fact, a close relative to *A. gigas*, the Asian arowana (*Scleropages formosus*) also a basal teleost belonging to the Osteoglossidae, presents two predicted Fads-like sequences recently deposited in GenBank KPP61181.1 and KPP71333.1 (not included in phylogenetic analysis due to their partial nature) annotated as FADS2-like and delta 6 desaturase-like respectively. However, no functional characterization these genes are yet available. Further studies are required to fully confirm the presence or absence of Fads1 in basal teleost lineages.

In conclusion, we herein demonstrate that *A. gigas* possess a *fads2* gene with all the typical features of front-end desaturases. Moreover, the functional assays of the *A. gigas* Fads2 in yeast confirmed that, like the majority of teleost Fads2, the *A. gigas* orthologue exhibited Δ6 and Δ8 desaturase activities. Along with the Fads2 from the Japanese eel (Wang et al., 2014), the herein reported *A. gigas* represents the most ancient representative of the Fads gene family being investigated within the teleost clade.

## Acknowledgements

This work was funded by the Projeto CAPES/Pró-Amazônia/Projeto Arapaima-Pirarucu/Universidade do Porto -CIIMAR. Fundação para a Ciência e a Tecnologia (FCT) PhD grant [grant number SFRH/BD/84238/2012] awarded to M. L-M. The access to the Institute of Aquaculture laboratories was funded by the European Union's Seventh Framework Programme (FP7/2007–2013) [grant number 262336] (AQUAEXCEL), Transnational Access [Project Number 0095/06/03/13].

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cbpb.2016.09.007.

## References

Abdul Hamid, N.K., Carmona-Antoñanzas, G., Monroig, Ó., Tocher, D.R., Turchini, G.M., Donald, J.A., 2016. Isolation and functional characterisation of a fads2 in rainbow trout (*Oncorhynchus mykiss*) with Δ5 desaturase activity. PLoS One 11 (3), e0150770.

Agaba, M., Tocher, D.R., Dickson, C.A., Dick, J.R., Teale, A.J., 2004. Zebrafish cDNA encoding multifunctional fatty acid elongase involved in production of eicosapentaenoic (20:5n-3) and docosahexaenoic (22:6n-3) acids. Mar. Biotechnol. (N.Y.) 6, 251–261.

Agaba, M.K., Tocher, D.R., Zheng, X., Dickson, C.A., Dick, J.R., Teale, A.J., 2005. Cloning and functional characterisation of polyunsaturated fatty acid elongases of marine and freshwater teleost fish. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 142, 342–352.

Andrade, J.I.A.d., Ono, E.A., de Menezes, G.C., Brasil, E.M., Roubach, R., Urbinati, E.C., Tavares-Dias, M., Marcon, J.L., Affonso, E.G., 2007. Influence of diets supplemented with vitamins C and E on pirarucu (*Arapaima gigas*) blood parameters. Comp. Biochem. Physiol. A Mol. Integr. Physiol. 146, 576–580.

Bauer, J.E., 1997. Fatty acid metabolism in domestic cats (*Felis catus*) and cheetahs (*Acinonyx jubatas*). Proc. Nutr. Soc. 56, 1013–1024.

Bell, M.V., Tocher, D.R., 2009. Biosynthesis of polyunsaturated fatty acids in aquatic ecosystems: general pathways and new directions. In: Kainz, M., Brett, M.T., Arts, M.T. (Eds.), Lipids in Aquatic Ecosystems. Springer New York, pp. 211–236.

Betts, M.J., Russell, R.B., 2003. Amino Acid Properties and Consequences of Substitutions, Bioinformatics for Geneticists. John Wiley & Sons, Ltd., pp. 289–316.

Bezerra, R.F., Soares, M.d.C.F., Santos, A.J.G., Maciel Carvalho, E.V.M., Coelho, L.C.B.B., 2014. Seasonality influence on biochemical and hematological indicators of stress and growth of *Pirarucu* (*Arapaima gigas*), an Amazonian air-breathing fish. Sci. World J. 2014, 6.

Brauner, C.J., Matey, V., Wilson, J.M., Bernier, N.J., Val, A.L., 2004. Transition in organ function during the evolution of air-breathing; insights from *Arapaima gigas*, an obligate air-breathing teleost from the Amazon. J. Exp. Biol. 207, 1433–1438.

Burdge, G.C., Calder, P.C., 2005. Conversion of alpha-linolenic acid to longer-chain polyunsaturated fatty acids in human adults. Reprod. Nutr. Dev. 45, 581–597.

Carmona-Antonanzas, G., Tocher, D.R., Taggart, J.B., Leaver, M.J., 2013. An evolutionary perspective on Elovl5 fatty acid elongase: comparison of northern pike and duplicated paralogs from Atlantic salmon. BMC Evol. Biol. 13, 85.

Castro, L.F.C., Monroig, Ó., Leaver, M.J., Wilson, J., Cunha, I., Tocher, D.R., 2012. Functional desaturase Fads1 (Δ5) and Fads2 (Δ6) orthologues evolved before the origin of jawed vertebrates. PLoS One 7, e31950.

Castro, L.F.C., Tocher, D.R., Monroig, O., 2016. Long-chain polyunsaturated fatty acid biosynthesis in chordates: insights into the evolution of fads and elovl gene repertoire. Prog. Lipid Res. 62, 25–40.

Cavero, B.A.S., Ituassú, D.R., Pereira-Filho, M., Roubach, R., Bordinhon, A.M., Fonseca, F.A.L., Ono, E.A., 2003. Uso de alimento vivo como dieta inicial no treinamento alimentar de juvenis de pirarucu. Pesq. Agrop. Brasileira 38, 1011–1015.

Fonseca-Madrigal, J., Navarro, J.C., Hontoria, F., Tocher, D.R., Martínez-Palacios, C.A., Monroig, Ó., 2014. Diversification of substrate specificities in teleostei Fads2: characterization of Δ4 and Δ6Δ5 desaturases of *Chirostoma estor*. J. Lipid Res. 55, 1408–1419.

Gonzalez, R.J., Brauner, C.J., Wang, Y.X., Richards, J.G., Patrick, M.L., Xi, W., Matey, V., Val, A.L., 2010. Impact of ontogenetic changes in branchial morphology on gill function in *Arapaima gigas*. Physiol. Biochem. Zool. 83, 322–332.

Goulding, M., 1980. Fishes and the Forest: Explorations in Amazonian Natural History. University of California Press.

Guillou, H., Zadravec, D., Martin, P.G., Jacobsson, A., 2010. The key roles of elongases and desaturases in mammalian fatty acid metabolism: insights from transgenic mice. Prog. Lipid Res. 49, 186–199.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.

Hastings, N., Agaba, M., Tocher, D.R., Leaver, M.J., Dick, J.R., Sargent, J.R., Teale, A.J., 2001. A vertebrate fatty acid desaturase with Δ5 and Δ6 activities. Proc. Natl. Acad. Sci. U. S. A. 98, 14304–14309.

Hastings, N., Agaba, M., Tocher, D., Zheng, X., Dickson, C., Dick, J., Teale, A., 2004. Molecular cloning and functional characterization of fatty acyl desaturase and elongase cDNAs involved in the production of eicosapentaenoic and docosahexaenoic acids from α-linolenic acid in Atlantic Salmon (*Salmo salar*). Mar. Biotechnol. 6, 463–474.

Ituassú, D.R., Pereira Filho, M., Roubach, R., Crescêncio, R., Cavero, B.A.S., Gandra, A.L., 2005. Níveis de proteína bruta para juvenis de pirarucu. Pesq. Agrop. Brasileira 40, 255–259.

Jump, D.B., Depner, C.M., Tripathy, S., 2012. Omega-3 fatty acid supplementation and cardiovascular disease: thematic review series: new lipid and lipoprotein targets for the treatment of cardiometabolic diseases. J. Lipid Res. 53, 2525–2545.

Kabeya, N., Yamamoto, Y., Cummins, S.F., Elizur, A., Yazawa, R., Takeuchi, Y., Haga, Y., Satoh, S., Yoshizaki, G., 2015. Polyunsaturated fatty acid metabolism in a marine teleost, nibe croaker *Nibea mitsukurii*: functional characterization of Fads2 desaturase and Elovl5 and Elovl4 elongases. Comp. Biochem. Physiol. Biochem. Mol. Biol. 188, 37–45.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9, 286–298.

Kuah, M.K., Jaya-Ram, A., Shu-Chien, A.C., 2015. The capacity for long-chain polyunsaturated fatty acid synthesis in a carnivorous vertebrate: functional characterisation and nutritional regulation of a Fads2 fatty acyl desaturase with Delta4 activity and an Elovl5 elongase in striped snakehead (*Channa striata*). Biochim. Biophys. Acta 1851, 248–260.

Li, Y., Monroig, O., Zhang, L., Wang, S., Zheng, X., Dick, J.R., You, C., Tocher, D.R., 2010. Vertebrate fatty acyl desaturase with Δ4 activity. Proc. Natl. Acad. Sci. 107, 16840–16845.

Lim, Z., Senger, T., Vrinten, P., 2014. Four amino acid residues influence the substrate chain-length and regioselectivity of *Siganus canaliculatus* Δ4 and Δ5/6 desaturases. Lipids 49, 357–367.

Liu, H., Guo, Z., Zheng, H., Wang, S., Wang, Y., Liu, W., Zhang, G., 2014. Functional characterization of a Δ5-like fatty acyl desaturase and its expression during early embryogenesis in the noble scallop *Chlamys nobilis* Reeve. Mol. Biol. Rep. 41, 7437–7445.

Los, D.A., Murata, N., 1998. Structure and expression of fatty acid desaturases. Biochim. Biophys. Acta Lipids Lipid Metab. 1394, 3–15.

Meesapyodsuk, D., Reed, D.W., Covello, P.S., Qiu, X., 2007. Primary structure, regioselectivity, and evolution of the membrane-bound fatty acid desaturases of *Claviceps purpurea*. J. Biol. Chem. 282, 20191–20199.

Miller, M.A., Schwartz, T., Pickett, B.E., He, S., Klem, E.B., Scheuermann, R.H., Passarotti, M., Kaufman, S., O'Leary, M.A., 2015. A RESTful API for access to phylogenetic tools via the CIPRES science gateway. Evol. Bioinformatics Online 11, 43–48.

Monroig, Ó., Zheng, X., Morais, S., Leaver, M.J., Taggart, J.B., Tocher, D.R., 2010. Multiple genes for functional 6 fatty acyl desaturases (Fad) in Atlantic salmon (*Salmo salar* L.): gene and cDNA characterization, functional expression, tissue distribution and nutritional regulation. Biochim. Biophys. Acta 1801, 1072–1081.

Monroig, Ó., Tocher, D.R., Navarro, J.C., 2011a. Long-chain polyunsaturated fatty acids in fish: recent advances on desaturases and elongases involved in their biosynthesis. In: Cruz-Suarez, L., Ricque-Marie, D., Tapia-Salazar, M., Nieto-López, M., Villarreal-Cavazos, D., Gamboa-Delgado, J., Hernández-Hernández, L. (Eds.), Décimo Primer Simposio Internacional de Nutrición Acuícola, San Nicolás de los Garza, N. L., México.

Monroig, Ó., Li, Y., Tocher, D.R., 2011b. Delta-8 desaturation activity varies among fatty acyl desaturases of teleost fish: high activity in delta-6 desaturases of marine species. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 159, 206–213.

Monroig, Ó., Wang, S., Zhang, L., You, C., Tocher, D.R., Li, Y., 2012. Elongation of long-chain fatty acids in rabbitfish *Siganus canaliculatus*: cloning, functional characterisation and tissue distribution of Elovl5- and Elovl4-like elongases. Aquaculture 350–353, 63–70.

Monroig, Ó., Tocher, D.R., Hontoria, F., Navarro, J.C., 2013. Functional characterisation of a Fads2 fatty acyl desaturase with Δ6/Δ8 activity and an Elovl5 with C16, C18 and C20 elongase activity in the anadromous teleost meagre (*Argyrosomus regius*). Aquaculture 412–413, 14–22.

Monroig, Ó., Hontoria, F., Varó, I., Tocher, D.R., Navarro, J.C., 2016a. Investigating the essential fatty acids in the common cuttlefish *Sepia officinalis* (Mollusca, Cephalopoda): molecular cloning and functional characterisation of fatty acyl desaturase and elongase. Aquaculture 450, 38–47.

Monroig, Ó., Lopes-Marques, M., Navarro, J.C., Hontoria, F., Ruivo, R., Santos, M.M., Venkatesh, B., Tocher, D.R., Castro, L.F., 2016b. Evolutionary functional elaboration of the Elovl2/5 gene family in chordates. Sci. Rep. 6, 20510.

Morais, S., Castanheira, F., Martinez-Rubio, L., Conceição, L.E.C., Tocher, D.R., 2012. Long chain polyunsaturated fatty acid synthesis in a marine vertebrate: ontogenetic and nutritional regulation of a fatty acyl desaturase with Δ4 activity. Biochim. Biophys. Acta 1821, 660–671.

Nelson, J.S., 1994. Fishes of the world. 3rd edition. John Wiley and Sons, Inc., New York 600 pp.

Oboh, A., Betancor, M.B., Tocher, D.R., Monroig, O., 2016. Biosynthesis of long-chain polyunsaturated fatty acids in the African catfish *Clarias gariepinus*: molecular cloning and functional characterisation of fatty acyl desaturase (fads2) and elongase (elovl2) cDNAs7. Aquaculture 462, 70–79.

Omasits, U., Ahrens, C.H., Müller, S., Wollscheid, B., 2014. Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics 30, 884–886.

Park, W.J., Kothapalli, K.S.D., Lawrence, P., Tyburczy, C., Brenna, J.T., 2009. An alternate pathway to long-chain polyunsaturates: the FADS2 gene product Δ8-desaturates 20:2n-6 and 20:3n-3. J. Lipid Res. 50, 1195–1202.

Park, H.G., Park, W.J., Kothapalli, K.S.D., Brenna, J.T., 2015. The fatty acid desaturase 2 (FADS2) gene product catalyzes Δ4 desaturation to yield n-3 docosahexaenoic acid and n-6 docosapentaenoic acid in human cells. FASEB J. 29, 3911–3919.

Pereira, S.L., Leonard, A.E., Mukerji, P., 2003. Recent advances in the study of fatty acid desaturases from animals and lower eukaryotes. Prostaglandins Leukot. Essent. Fat. Acids 68, 97–106.

Perica, M.M., Delaš, I., 2011. Essential fatty acids and psychiatric disorders. Nutr. Clin. Pract. 26, 409–425.

Psota, T.L., Gebauer, S.K., Kris-Etherton, P., 2006. Dietary omega-3 fatty acid intake and cardiovascular risk. Am. J. Cardiol. 98, 3–18.

Ribeiro, R.A., Ozório, R.O.d.A., Batista, S.M.G., Pereira-Filho, M., Ono, E.A., Roubach, R., 2011. Use of spray-dried blood meal as an alternative protein source in *Pirarucu* (*Arapaima gigas*) diets. J. Appl. Aquac. 23, 238–249.

Rivers, J.P.W., Sinclair, A.J., Crawford, M.A., 1975. Inability of the cat to desaturate essential fatty acids. Nature 258, 171–173.

Robinson, L., Mazurak, V., 2013. N-3 polyunsaturated fatty acids: relationship to inflammation in healthy adults and adults exhibiting features of metabolic syndrome. Lipids 48, 319–332.

Rose, T.M., Henikoff, J.G., Henikoff, S., 2003. CODEHOP (COnsensus-DEgenerate hybrid oligonucleotide primer) PCR primer design. Nucleic Acids Res. 31, 3763–3766.

Shanklin, J., Whittle, E., Fox, B.G., 1994. Eight histidine residues are catalytically essential in a membrane-associated iron enzyme, stearoyl-CoA desaturase, and are conserved in alkane hydroxylase and xylene monooxygenase. Biochemistry (Mosc) 33, 12787–12794.

Tanomman, S., Ketudat-Cairns, M., Jangprai, A., Boonanuntanasarn, S., 2013. Characterization of fatty acid delta-6 desaturase gene in Nile tilapia and heterogenous expression in *Saccharomyces cerevisiae*. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 166, 148–156.

Tocher, D.R., 2003. Metabolism and functions of lipids and fatty acids in teleost fish. Rev. Fish. Sci. 11, 107–184.

Tocher, D.R., 2010. Fatty acid requirements in ontogeny of marine and freshwater fish. Aquac. Res. 41, 717–732.

Tocher, D.R., Leaver, M.J., Hodgson, P.A., 1998. Recent advances in the biochemistry and molecular biology of fatty acyl desaturases. Prog. Lipid Res. 37, 73–117.

Torres, F.G., Malásquez, M., Troncoso, O.P., 2015. Impact and fracture analysis of fish scales from *Arapaima gigas*. Mater. Sci. Eng. C 51, 153–157.

Tsirigos, K.D., Peters, C., Shu, N., Käll, L., Elofsson, A., 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 43, W401–W407.

Turchini, G.M., Torstensen, B.E., Ng, W.-K., 2009. Fish oil replacement in finfish nutrition. Rev. Aquac. 1, 10–57.

Veríssimo, J., 1895. A pesca na Amazônia. Livraria classica de Alves.

Wall, R., Ross, R.P., Fitzgerald, G.F., Stanton, C., 2010. Fatty acids from fish: the anti-inflammatory potential of long-chain omega-3 fatty acids. Nutr. Rev. 68, 280–289.

Wang, S., Monroig, Ó., Tang, G., Zhang, L., You, C., Tocher, D.R., Li, Y., 2014. Investigating long-chain polyunsaturated fatty acid biosynthesis in teleost fish: functional characterization of fatty acyl desaturase (Fads2) and Elovl5 elongase in the catadromous species, Japanese eel *Anguilla japonica*. Aquaculture 434, 57–65.

Xie, D., Chen, F., Lin, S., Wang, S., You, C., Monroig, Ó., Tocher, D.R., Li, Y., 2014. Cloning, functional characterization and nutritional regulation of Δ6 fatty acyl desaturase in the herbivorous euryhaline teleost *Scatophagus Argus*. PLoS One 9, e90200.

Zheng, X., Seiliez, I., Hastings, N., Tocher, D.R., Panserat, S., Dickson, C.A., Bergot, P., Teale, A.J., 2004. Characterization and comparison of fatty acyl Δ6 desaturase cDNAs from freshwater and marine teleost fish species. Comp. Biochem. Physiol. Biochem. Mol. Biol. 139, 269–279.

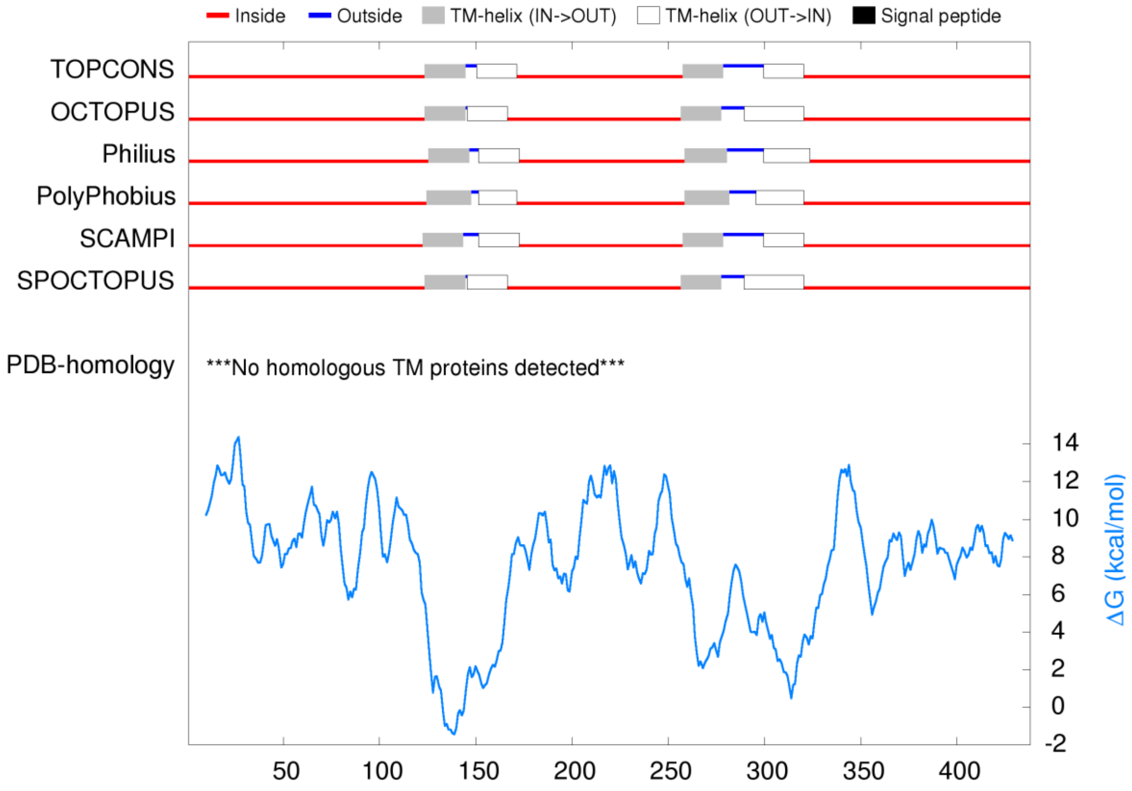
Zheng, X., Tocher, D.R., Dickson, C.A., Bell, J.G., Teale, A.J., 2005. Highly unsaturated fatty acid synthesis in vertebrates: new insights with the cloning and characterization of a delta6 desaturase of Atlantic salmon. Lipids 40, 13–24.

Zheng, X., Ding, Z., Xu, Y., Monroig, O., Morais, S., Tocher, D.R., 2009. Physiological roles of fatty acyl desaturases and elongases in marine fish: characterisation of cDNAs of fatty acyl Δ6 desaturase and elovl5 elongase of cobia (*Rachycentron canadum*). Aquaculture 290, 122–131.

## SUPPLEMENTARY MATERIAL

2D topology prediction results

| Method | TM- helix position starting from 1 | | | |
|---|---|---|---|---|
| **TOPCONS** | TM1: 124-145, | TM2: 151-172, | TM3: 258-279, | TM4: 300-321 |
| **OCTOPUS** | TM1: 124-145, | TM2: 146-167, | TM3: 257-278, | TM4: 290-321 |
| **Philius** | TM1: 126-147, | TM2: 152-173, | TM3: 259-281, | TM4: 300-324 |
| **PolyPhobius** | TM1: 125-148, | TM2: 152-172, | TM3: 259-282, | TM4: 296-321 |
| **SCAMPI** | TM1: 123-144, | TM2: 152-173, | TM3: 258-279, | TM4: 300-321 |
| **SPOCTOPUS** | TM1: 124-145, | TM2: 146-167, | TM3: 257-278, | TM4: 290-321 |



## IV.3 Gene duplication and loss underscore the vertebrate efficiency in completing long-chain polyunsaturated fatty acids biosynthesis.

MÓNICA LOPES-MARQUES; NAOKI KABEYA; YIAN QIAN; RAQUEL RUIVO, MIGUEL M. SANTOS, DOUGLAS R. TOCHER, BYRAPPA VENKATESH, L. FILIPE. C. CASTRO, ÓSCAR. MONROIG

**IN PREPARATION**

# GENE DUPLICATION AND LOSS UNDERSCORE THE VERTEBRATE EFFICIENCY IN COMPLETING LONG-CHAIN POLYUNSATURATED FATTY ACIDS BIOSYNTHESIS

Mónica Lopes-Marques[1,2], Naoki Kabeya[3,4], Yu Qian[3], Raquel Ruivo[1], Miguel Santos[1,5], Douglas R. Tocher[3], Byrappa Venkatesh[6], L. Filipe C. Castro[1,5]*, and Oscar Monroig[3]*

1-CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, U. Porto – University of Porto, Porto, Portugal

2-ICBAS - Institute of Biomedical Sciences Abel Salazar, U. Porto - University of Porto, Portugal

3-Institute of Aquaculture, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK

4-Department of Aquatic Bioscience, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1, Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

5-Department of Biology, Faculty of Sciences, U. Porto - University of Porto, Portugal

6-Comparative Genomics Laboratory, Institute of Molecular and Cell Biology, A*STAR (Agency for Science, Technology and Research), Biopolis, Singapore

*Corresponding authors

**Keywords:** LC-PUFA biosynthesis, Fatty acid desaturase (*fads*), basal actinopterygii

## 1. INTRODUCTION

Assessing the impact of genome/gene duplication and gene loss in shaping metabolic pathways is critical to understand how metabolic pathways are attuned in several species (Shimeld *et al.*, 2000; Cañestro, 2012; Chen *et al.*, 2013). Presently it is generally accepted that two rounds of whole genome duplication occurred in early vertebrate evolution approximately 500 MYA (Ohno; Putnam *et al.*, 2008). Additional events of genome duplications have been also documented in the teleost ancestor (3R WGD) (Jaillon *et al.*, 2004) as well as, lineage or species specific duplications in salmonids (Moghadam *et al.*, 2011), ray-finned paddle fish (Crow *et al.*, 2012), African clawed frog (Session *et al.*, 2016), and the red viscacha rat (Gallardo *et al.*, 1999; Gallardo *et al.*, 2004). Studies focusing on key phylogenetic lineages within the chordates have revealed the impact of these events on the gain and/or loss of several developmental, morphological and physiological features (Braasch *et al.*, 2014; Castro *et al.*, 2014; Brunet *et al.*, 2016; Monroig *et al.*, 2016).

The biosynthesis of Long-Chain Polyunsaturated Fatty Acids (LC-PUFAS) in vertebrates clearly exemplifies how gene repertoire and enzymatic capabilities significantly impact the extent to which each species can complete LC-PUFAS biosynthesis pathway (Carmona-Antonanzas *et al.*, 2013; Castro *et al.*, 2016; Monroig *et al.*, 2016). LC-PUFAS such as arachidonic acid (ARA, 20:4n-6), eicosapentaenoic acid (EPA, 20:5n-3) and docosahexaenoic acid (DHA, 22:6n-3) are key biomolecules with significant roles in physiological processes such as reproduction (Robinson *et al.*, 2013) inflammatory response (Wall *et al.*, 2010) neural development (Perica *et al.*, 2011) bio-membrane composition and energy storage (Tocher, 2003). In animals, LC-PUFAS are obtained through diet and/or may be endogenously synthesized from essential dietary fatty acids (EFA) such as linoleic and α-linoleic acid (LA-18:2n-6; ALA-18:3n-3). Biosynthesis of LC-PUFAS comprehends a series of consecutive reactions of elongation e desaturation backed by elongases (Elovl) and desaturases (Fads) enzymes (Tocher, 2003; Monroig *et al.*, 2011b). This metabolic pathway can be split in to two major steps: initial transformation of EFA with the conjoint action of both *fads2* and *elovl5* to obtain eicosatetraenoic acid or dihomo-gamma-linolenic acid (ETA- 20:4(n-3); DGLA-20:4(n-6)); and further elongation and desaturation of these products promoted by *elovl2* and *fads1* to obtain tetracosahexaenoic and tetracosapentaenoic acid ((24:6(n-3) (24:5n-6)) (Tocher, 2003; Schmitz *et al.*, 2008; Monroig *et al.*, 2011b). Consequently, a full set of elongase enzymes (Elovl2 and Elovl5) as well as, desaturase enzymes (Fads1 and Fads2) are assumed crucial to complete this biosynthetic pathway.

Although the diversification of *Elovl2* and *Elovl5* in vertebrates has been attributed to the 2R WGD (Monroig *et al.*, 2016) the origin and distribution of *Fads1* and *Fads2* still remains to be fully clarified. The identification of *Fads1* and *Fads2* orthologues in the cartilaginous fish *Scyliorhinus canicula* provides a solid indication on the timing of the expansion of these genes in the pre-ganthostome ancestor (Castro *et al.*, 2012); then followed by the loss of *Fads1* in teleosts (Castro *et al.*, 2012; Castro *et al.*, 2016). Interestingly besides *Fads1, Elovl2* has also been reported to be lost in many teleosts (Morais *et al.*, 2009; Monroig *et al.*, 2016), with the exception of *Danio rerio* (Monroig *et al.*, 2009)*, Salmo salar* (Morais *et al.*, 2009)*,* and *Oncorhynchus mykiss* (Gregory *et al.*, 2014). Therefore, variable *Elovl* and *Fads* gene repertoires and substrates preferences can be found in several vertebrate lineages. For instance, humans present three *Fads* genes: *Fads1, Fads2,* and *Fads3*, organized in a gene cluster with the *Fads3* gene function so far to be clarified (Marquardt *et al.*, 2000; Blanchard *et al.*, 2011). While mammalian desaturase enzymes display essentially a single preferred desaturation activity where *Fads1* is a Δ5 desaturase and *Fads2* is a Δ6 desaturase with additional Δ4 activity in some mammals (Park *et al.*, 2009a; Park *et al.*, 2015); previous research has uncovered functional plasticity of fatty acid desaturases in teleost fish. Despite *Fads2* being the unique orthologous desaturase found so far in teleostei fish, a wide spectrum of alternative substrate preferences has been found: mono-functional Δ6 desaturase in *Thunnus thynnus* (Morais *et al.*, 2011)*, Dicentrarchus labrax* (González-Rovira *et al.*, 2009; Santigosa *et al.*, 2011) and *Lates calcarifer* (Mohd-Yusof *et al.*, 2010); Δ5 desaturase in *Salmo salar* (Hastings *et al.*, 2004); *fads2* with dual desaturation activity: Δ4/Δ5 *Chirostoma estor* (Fonseca-Madrigal *et al.*, 2014), *Solea senegalensis* (Morais *et al.*, 2012)*, Channa striata* (Kuah *et al.*, 2015); Δ5/Δ6: *Oreochromis niloticus* (Tanomman *et al.*, 2013); Δ6/Δ8: *Nibea mitsukurii* (Kabeya *et al.*, 2015), *Argyrosomus regius* (Monroig *et al.*, 2013), *O. mykiss* (Seiliez *et al.*, 2001; Zheng *et al.*, 2004; Monroig *et al.*, 2011a) *Anguilla japonica* (Wang *et al.*, 2014); Δ4/Δ5: *Channa striata* (Kuah *et al.*, 2015), *Chirostoma estor* (Fads2a) (Fonseca-

Madrigal *et al*., 2014), *Solea senegalensis* (Morais *et al*., 2012), and lastly *fads2* with triple desaturation capacity Δ5/Δ6/Δ8 found in *D. rerio* (Hastings *et al*., 2001; Monroig *et al*., 2011a) *Siganus canaliculatus* (Li *et al*., 2010; Monroig *et al*., 2011a) *C. estor (Fads2b)* (Fonseca-Madrigal *et al*., 2014); and Δ4/Δ5/Δ8 in *S. canaliculatus* (Fad2) (Li *et al*., 2010; Monroig *et al*., 2011a). Therefore, the sole presence or absence of a complete gene set of *Fads* is insufficient to infer to what degree a certain species can convert dietary EFA to LC-PUFA.

To fully understand the impact of gene duplication and loss in the LC-PUFA biosynthesis we have isolated and functionally characterized *Fads* from species placed in key phylogenetic positions namely, the basal cyclostome *Lethenteron japonicum* (Japanese lamprey) that diverged from the from the gnathostome ancestor approximately at the time of the 2R WGD (Kuraku *et al*., 2009; Smith *et al*., 2015), and from four actinopterygii species, two that diverged before the teleost specific 3R WGD namely the polypteriforme *Polypterus senegalus* (bichir) and holostei *Lepisosteus oculatus* (spotted gar) (Amores *et al*., 2011), and two that diverged after the teleost specific 3RWGD the elopomorpha *Anguilla japonica* (Japanese ell) and the osteoglossomorpha *Pantodon buchholzi* (African butterfly fish) (Betancur-R. R *et al*., 2013).

## 2. MATERIALS AND METHODS

### 2.1 SEQUENCE COLLECTION AND ASSEMBLY

Fads amino acid sequences were recovered from the available databases Ensembl and GenBank. *Scleropages formosus* presented a 3´partial fads sequence XP_018598908.1, this sequence was completed by performing blastn searches in *S. formosus* transcriptome SRA reads (SRX1668426/27/28/29/30/31/32). *Gnathonemus petersii* and *Osteoglossum bicirrhosum fads-like* were assembled from genomic SRX2235995, SRX2235994 in Geneious V 7.1.9 using as reference the previously curated *S. formosus fads.*

Regarding the *Fads* isolated in this work initial tblastn searches using as query *S. canicula Fads1* (AEY94454.1) *and Fads2* (AEY94455.1) were performed in the Japanese lamprey genome project available at (http://jlampreygenome.imcb.a-star.edu.sg/) for *L. japonicum;* NCBI sequence read archives (SRX796491, SRX732875) for *P. senegalus;* (SRX666400) for *P. buchholzi* and the genomic assembly of *A. japonica* (KI1307852) also available at NCBI. The resulting hits were downloaded and assembled into predicted full ORFs (open reading frames) using as reference the corresponding bait sequences. In the case of *L. oculatus Fads1* the available sequence (XM_015338726.1) represents a truncated isoform, we further searched *L. oculatus* SRA archives with the resulting hits retrieved and assembled to reveal a second non-truncated isoform of *fads1*. All final assembled/predicted *fads* sequences were further used as reference for primer design.

### 2.2 PHYLOGENETIC ANALYSIS

A total of 79 Fads amino acid were aligned with MAFFT v7.306 (Katoh *et al.*, 2008) and the best alignment method was determined automatically resulting in L-INS-i method (Katoh *et al.*, 2005). Columns containing 90% gaps were stripped from sequence alignment leaving a total of 452 sites for phylogenetic analysis. Sequence alignment was then submitted PhyML v3.0 server (Guindon *et al.*, 2010) for maximum likelihood phylogenetic analysis. The evolutionary model was automatically selected by the smart model selection SMS resulting in LG+G+I, and branch support was calculated using Abayes (Anisimova *et al.*, 2011). The resulting tree was visualized in Fig Tree V1.3.1 available at http://tree.bio.ed.ac.uk/software/figtree/ and rooted with invertebrate desaturase-like sequences.

### 2.3 *FADS* GENE ISOLATION AND CLONING INTO YEAST EXPRESSION VECTOR

The total RNA was extracted from *L. japonicum, P. senegalus*, *L. oculatus,* using an Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, UK) and *A. japonica* using ISOGEN (NIPPON GENE CO., LTD., Tokyo, Japan) following the manufacturer's recommendations. Extracted RNA was quantified and the integrity evaluated. Complementary DNA was then synthesized using the iScript cDNA Synthesis Kit (Bio-Rad) for *L. japonicum, P. senegalus*, *L. oculatus* and for *A. japonica* Ready-To-Go You-Prime First-Strand Beads (GE Healthcare Life Sciences, Chicago, IL, USA) according to manufacturer's guidelines. 3´RACE ready cDNA (Rapid Amplification of cDNA Ends) was prepared for *L. oculatus.* For *Fads* gene isolation and cloning into yeast expression vector gene specific primers containing the appropriate restriction sites were designed on the previously recovered or assembled sequences. *Fads* genes were then isolated in each specie using these primers and corresponding cDNA via polymerase chain reactions (PCR) (see Table 1 for primer details and PCR conditions). Resulting PCR products were analyzed by electrophoresis in 1% agarose gel the target PCR products were excised and purified. Each PCR product was digested with the corresponding restriction enzymes and ligated into the yeast expression vector pYES2 (Thermo Fisher Scientific, Waltham, MA, USA) using T4 DNA ligase (Promega). Finally, all pYES2 clones were confirmed by sequencing (GATC Biotech Constance, Germany).

**Table 1.** Primer sets, corresponding PCR conditions. Details regarding *L. japonicum* are still to be provided by co-author B. Venkatesh

| | PCR details | | Primer sequence | Initial denaturation | Cycles | Denaturation | TM | Extension | Final extension |
|---|---|---|---|---|---|---|---|---|---|
| ***Lethenteron japonicum*** | - | ***Fads1*** | **FW :**CCC**AAGCTT**CACCATGGGACGCGGCGA<br>**RV:** CCC**TCTAGA**TCACTTGTGCAGGTAAGCGTC | - | - | - | - | - | - |
| | - | ***Fads2*** | **FW**: CCC**AAGCTT**ACCATGGCTGGAACAGCATCG<br>**RV**: CCC**TCTAGA**TCACCGCTGGAGGTAGGCAT | - | - | - | - | - | - |
| ***Polypterus senegalus*** | Phusion Flash High-Fidelity PCR Master Mix | ***Fads1*** | **FW:** CCC**GGTACC**ATGGAGGATGAAACAAAAGATAAAA<br>**RV:** CCC**TCTAGA**TCACTTATGCAGGTAGGCGTC | 98ºC /10s | 45 | 98ºC /1s | 61ºC/5s | 72ºC/21s | 72ºC/1min |
| | | ***Fads2*** | **FW:** CCC**GGTACC**CCTAAAATGGGGAAAGGTGG<br>**RV:** CCC**TCTAGA**GTTTCTCTCTTTCTTACTTGTTAAG | 98ºC /10s | 35 | 98ºC /1s | 61ºC/5s | 72ºC/21s | 72ºC/1min |
| ***Lepisosteus oculatus*** | Phusion Flash High-Fidelity PCR Master Mix | ***Fads1 Long*** | **FW**: CCC**GGATCC**AGGATGGGCGCAGGCGCAGA<br>**RV**: CCG**TCTAGA**TCACCTGTGCAGGTAGGCATCAAGC | 98ºC /10s | 40 | 98ºC /1s | 69ºc/5s | 72ºC/20s | 72ºC/1min |
| | Phusion Flash High-Fidelity PCR Master Mix+ 3%DMSO | ***Fads2*** | **FW:** CCC**GGTACC**ACAATGGGTGGGGGGGGCCAGC<br>**RV:** CCC**TCTAGA**CCTATTTGTGGAGGTAGGCATCCA | 98ºC /30s | 40 | 98ºC /1s | 68ºc/5s | 72ºC/20s | 72ºC/30s |
| ***Pantodon buchholzi*** | Phusion Flash High-Fidelity PCR Master Mix | ***Fads2a*** | **FW:** CCC**GGTACC**ATGGGAGGCGGTGGGCAGC<br>**RV:** CCC**TCTAGA**TCACTTATGCAGGTAGGCATCCAG | 98ºC /10s | 35 | 98ºC /1s | 65ºC/5s | 72ºC/20s | 72ºC/1min |
| | | ***Fads2b*** | **FW:** CCC**GGATCC**AATATGGGTGGTGGAGGACAGC<br>**RV**: CCC**CTCGAG**TTACTTGTGAAGGTACGCATCCAG | 98ºC /10s | 35 | 98ºC /1s | 65ºC/5s | 72ºC/20s | 72ºC/1min |
| ***Anguilla japonica*** | *Pfu* DNA polymerase | ***Fads1*** | **FW:**CCC**AAGCTT**AACATGAGCGCAGCAGAGAAG<br>**RV:**CCG**TCTAGA**TCATTTGTGCAAGTAAGCATCCATC | 95ºC/2min | 35 | 95ºC /30s | 60ºC/30s | 72ºC/3min | 72ºC/5min |

## 2.4 YEAST EXPRESSION ASSAYS AND FATTY ACID ANALYSIS

Transformation with pYes2 expression vector and culture of yeast *Saccharomyces cerevisiae* was carried out by the method described previously (Hastings *et al*., 2001; Lopes-Marques *et al*., 2017). The resulting transgenic yeast expressing each *Fads* genes were grown in the presence of PUFA including Δ6 (18:3n-3 and 18:2n-6), Δ8 (20:2n-6 and 20:3n-3), Δ5 (20:4n-3; 20:3n-6), Δ4 (22:5n-3 and 22:4n-6) desaturase substrates.

The final concentrations of PUFA substrate were 0.5 mM ($C_{18}$), 0.75 mM ($C_{20}$) and 1.0 mM ($C_{22}$). All FA substrates (98–99% pure) except for 18:4n-3 and 20:4n-3 were purchased from Nu-Chek Prep, Inc. (Elysian, MN, USA). 18:4n-3 and 20:4n-3 were obtained from Sigma-Aldrich (St Louis, MO, USA) and Cayman Chemical (Ann Arbor, MI, USA), respectively. After 48 h incubation at 30ºC with shaking, yeast cells were collected and fatty acid methyl esters (FAME) were prepared by the method described previously (Hastings *et al*., 2001). Subsequently, the FAME samples were quantified using Fisons GC-8160 gas chromatograph (Thermo Fisher Scientific) equipped with a 60 m x 0.32 mm i.d. x 0.25 μm ZB-wax column (Phenomenex, Cheshire, UK) and flame ionisation detector. The obtained FAME peaks were identified by comparing the retention time of each with that of FAME standard. FA conversion efficiencies from exogenously added PUFA substrates were calculated by the proportion of substrate FA converted to a desaturated product as (product area/(product area + substrate area)) × 100.

## 3 RESULTS

### 3.1 PHYLOGENETIC ANALYSIS

To determine the orthology of isolated *fads-like* sequences from the jawless vertebrate the Japanese lamprey and actinopterygii species a Bayesian phylogenetic analysis was conducted containing 79 sequences, of which several are well known and functionally characterized. The resulting phylogenetic tree presents two well supported monophyletic clades; the first holding all the Fads1 sequences and the second containing the Fads2 sequences, both clades are out-grouped by invertebrate Fads sequences (Fig. 1). Interestingly, we find that the Japanese lamprey, bichir, spotted gar and Japanese ell present sequences that strongly group within the Fads1 clade together with the tetrapod and chondrichthyes Fads1 sequences.

When observing the Fads2 clade we find that Japanese lamprey Fads2-like is again basal to the Fads2 clade. Within the Fads2 clade we find all the actinopterygii Fads2 are placed together in a single clade contiguous to a sister clade containing the sarcopterigyii and chondrichthyes Fads2 sequences. Internal topology of the actinopterygii Fads2 clade reflects evolutionary history of ray-finned-fishes, were the lineages that diverged before the 3RWGD namely: polypteriformes (*P. senegalus*) and holostei (*L. oculatus*), are placed at the base of the clade, followed by post 3R WGD lineages elopomorpha (*A. japonica*) and the osteoglossomorpha (*P. buchholzi*, *Arapaima gigas; S. formosus; Osteoglossum bicirrhosum, Gnathonemus petersii*) and finally all the remaining teleostei. In osteoglossmorpha clade we observe all species present two fads2 sequences (with the exception of *A. gigas*) which are distributed equally among two well supported clades (0.9). The splitting of each species Fads2 duplicates into two clades and branching is indicative that these sequences may correspond to retained 3RWGD paralogs (Fig. 1).

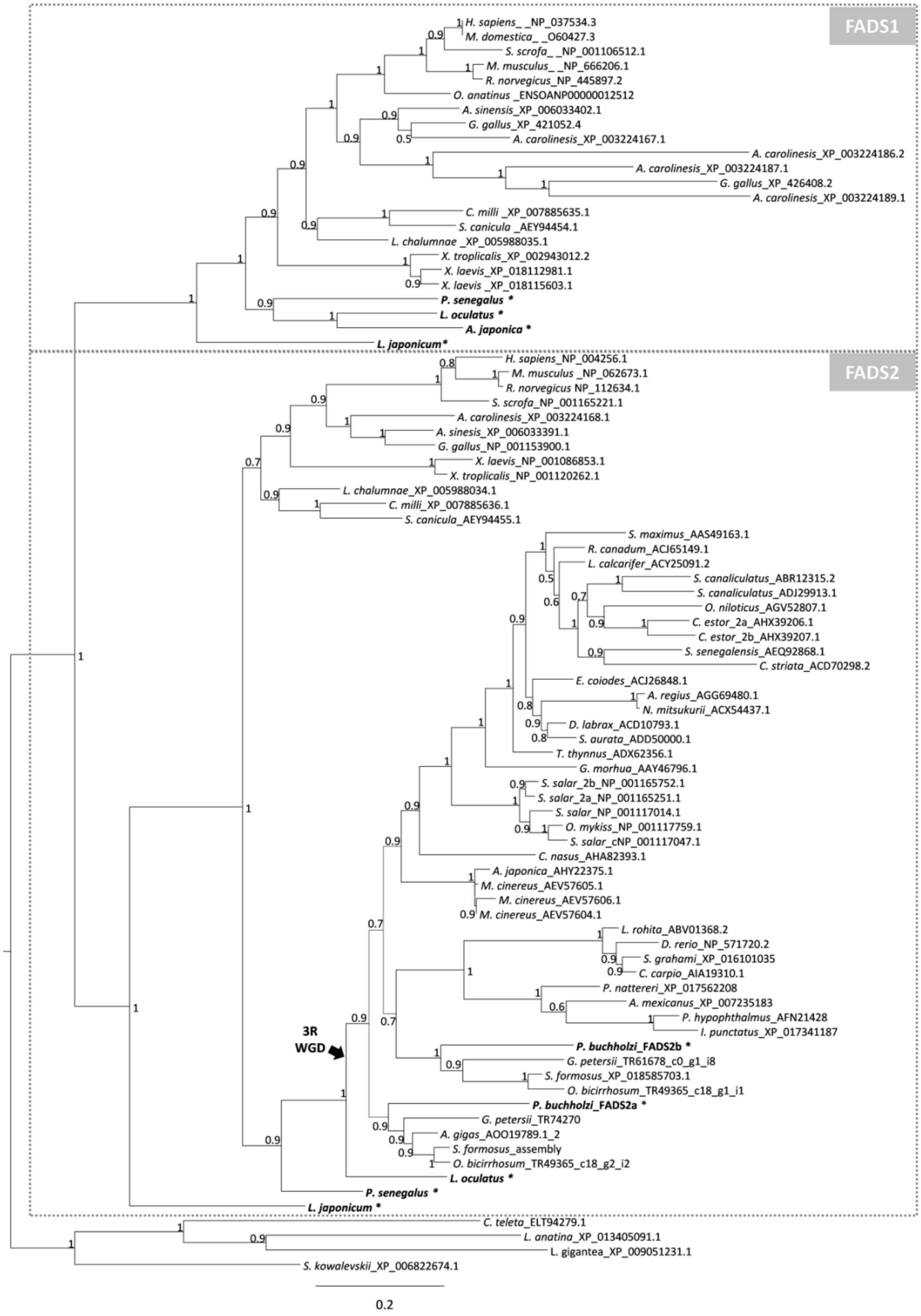


**Figure1:** Bayesian phylogenetic analysis of FADS1 and FADS2 amino acid sequences, values at nodes indicate posterior probabilities, * indicates FADS isolated and functionally analyzed in this work. Black arrow (3R WGD) approximately indicates the timing of the teleost duplication. Accession numberes are indicated.

### 3.2 Sequence analysis

All 9 fads-like sequences isolated from *A. japonica*, *P. buchholzi*, *L. oculatus*, *P. senegalus*, *L. japonicum*, were submitted to PFam to validate *fads-like* profile, revealing a *fads1-like* or *fads-like2* like profile. Initial sequence analysis revealed that *L. oculatus fads1* and *fads2* gene annotations available at NCBI and Ensemble (*fads1*- XM_015338726.1, ENSLOCG00000007048, *fads2*- ENSLOCG00000007031.1) correspond to poor gene predictions due to low genome coverage in the respective regions. In the case of *fads1-like* the predicted gene omitted the 5´region of exon 2 and 3´region of the last exon, further analysis of genomic sequence in NCBI (NC_023205.1) revealed a premature stop codon in the 3´region of the last exon. Regarding the predicted *L. oculatus fads2* this annotation was overall poor due to several assembly gaps. The isolated ORF *L. oculatus fads1-like* and *fads2-like* fully covered the poorly predicted regions. Furthermore, the *fads1-like* clone did not present the premature stop codon; additionally both clones were confirmed by *L. oculatus* transcriptomic SRA reads. The isolated *fads*-like sequences were further aligned with fully characterized fads sequences (Fig. 2).

Sequence alignment revealed that all sequences presented the signature motifs characteristic of FADS, namely the heme binding motif (HPGG) and three histidine boxes HXXXH, HXXHH, and QXXHH, suggested to participate in Fe binding in the active site of the enzyme (Los *et al.*, 1998; Pereira *et al.*, 2003). In all isolated sequences, we find that the heme motif is as well as, all three histidine boxes are highly conserved. Next, we analyzed the amino acid residues proposed to be critical in switching from Δ6 desaturase activity to Δ5 and vice versa (Watanabe *et al.*, 2016). These residues were identified in rat desaturase by replacing *Fads2* Δ6 desaturase residues in the following positions [**S**X**N**XXXX**R**X …**SL**X… **WQ**X…**V**] (Ser209, Asn211, Arg216, Ser235, Leu236, Trp244, Gln245, and Val344 - coordinates of Human Fads2 NP_004256.1) with residues from Δ5 desaturase [**P**X**S**XXXX**M**X … **M-**X … **VL**X…**P**] obtaining Δ5 desaturation (Watanabe *et al.*, 2016) (Fig. 2 blue boxes). In the sequence alignment, we observe that the subset of sequences from *A. japonica, L. oculatus, P. senegalus* and *L. japonicum* that returned a *fads1-like* profile from pFAM searches present the majority of the residues typical of Δ5 desaturation with two substitutions M>L and V>T corresponding to conservative replacements between residues with similar biochemical properties.

Interestingly, *P. buchholzi Fads2-likeb* (PBU) and Asian arowana *Fads2* (AGI) present residues typical of Δ5 desaturation indicated by the black arrows. Additionally, a second region also proposed to be involved in regioselectivity (Lim *et al.*, 2014) (Fig. 2 yellow box) revealed that all *fads1*-like sequences preserve a fully conserved signature - FQWI, while the *fads2–like* sequences showed to be more variable in this region presenting the following pattern FXXQ.

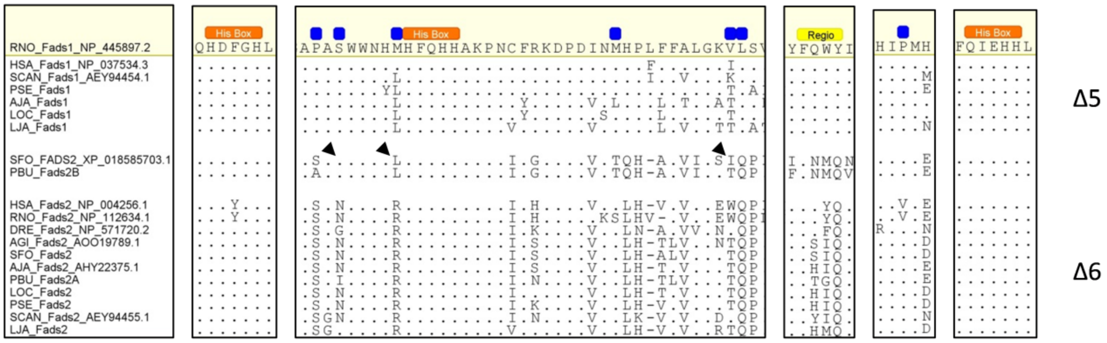


**Figure 2:** Sequence alignment of FADS1 and FADS2 amino acid sequences. Orange boxes correspond to the conserved histidine boxes, yellow box indicates residues proposed to be involved in regioselectivity (Lim *et al.*, 2014), blue boxes indicate residues replaced in rat *fads2* Δ6 desaturase to obtain Δ5 activity (Watanabe *et al.*, 2016). RNO- *R. norvegicus;* HSA- *H. sapiens;* SCAN- *S. canicula*; PSE- *P. senegalus;* AJA- *A. japonica;* LOC- *L. oculatus;* LJA- *L. japonicum;* SFO- *S. formosus* PBU-*P. buchholzi*; DRE-*D. rerio;* AGI-*A. gigas*.

### 3.3 FUNCTIONAL ANALYSIS SHOWS PLASTICITY WITH TELEOST FADS

The expression of the isolated Fads ORFs in yeast grown in media supplemented potential desaturase PUFAS substrates Δ6 (18:3n-3 and 18:2n-6), Δ8 (20:2n-6 and 20:3n-3), Δ5 (20:4n-3 and 20:3n-6) and Δ4 (22:5n-3 and 22:4n-6) allow us to functionally characterize each FADS enzyme and determine the preferred desaturation activity. The FA profile of yeast transformed with empty vector was determined as control and showed the following products: yeast endogenous FA (16:0), (16:1n-7), (18:0) and (18:1n-9), and corresponding supplemented PUFA substrate confirming that the yeast endogenous enzymes were not active on the exogenously added PUFA substrates (Agaba *et al.*, 2005).

Functional characterization showed that Fads1-like ORFs isolated from Japanese lamprey, bichir, spotted gar and Japanese eel coded for enzymes that presented Δ5 desaturation activity by desaturating 20:4n-3 and 20:3n-6 substrates, with a preference towards omega-3 PUFAS (Table 2).

Yeast containing Fads2-like ORFs from Japanese lamprey, bichir, spotted gar and African butterfly fish (fads2a-like) expressed desaturase enzymes with Δ6/Δ8 desaturation activities with the exception of the African butterfly fish Fads2b-like that presented Δ5 desaturation activity. Conversion rates indicate that in all cases Δ6 desaturase activity is the most predominant (Table 2). Regarding Δ8 desaturase activity we observe that Japanese lamprey Fads2 presents no detectable desaturation of the Δ8 substrate 20:2n-6 substrate (Table 2). Finally, no FA products resulting from Δ4 desaturase activity were detected in all functional assays.

**Table 2:** Functional characterization of the L. *japonicum* (Lja); *P. senegalus* (Pse); *L. oculatus* (Loc); *A. japonica* (Aja); *P. buchholzi* (Pbu); desaturase enzymes. The conversions were calculated according to the formula (all product areas/ (all products areas+substrate area)) ×100 and .n.d indicates not detected. **a**- ( Wang *et al* 2014)

| | | **% Conversion** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FA Substrate** | **FA Product** | **Lja** Fads1 | **Lja** Fads2 | **Loc** Fads1 | **Loc** Fads2 | **Pse** Fads1 | **Pse** Fads2 | **Aja** Fads1 | **Aja** [a] Fads2 | **Pbu** Fads2A | **Pbu** Fads2B | **Activity** |
| 18:3n-3 | 18:4n-3 | n.d | 6.6 | n.d | 32.4 | n.d | 37.1 | n.d | 64.3 | 77.4 | n.d | Δ6 |
| 18:2n-6 | 18:3n-6 | n.d | 2.0 | n.d | 15.6 | n.d | 20.6 | n.d | 20.7 | 42.7 | n.d | Δ6 |
| 20:3n-3 | 20:4n-3 | n.d | 0.7 | n.d | 4.1 | n.d | 11.0 | n.d | 6.0 | 18.4 | n.d | Δ8 |
| 20:2n-6 | 20:3n-6 | n.d | n.d | n.d | 1.5 | n.d | 3.6 | n.d | 5.4 | 7.0 | n.d | Δ8 |
| 20:4n-3 | 20:5n-3 | 6.0 | n.d | 3.0 | n.d | 56.1 | n.d | 58.1 | n.d | n.d | 14.4 | Δ5 |
| 20:3n-6 | 20:4n-6 | 5.5 | n.d | 2.9 | n.d | 48.3 | n.d | 33.2 | n.d | n.d | 11.7 | Δ5 |
| 22:5n-3 | 22:6n-3 | n.d | n.d | n.d | n.d | n.d | n.d | n.d | n.d | n.d | n.d | Δ4 |
| 22:4n-6 | 22:5n-6 | n.d | n.d | n.d | n.d | n.d | n.d | n.d | n.d | n.d | n.d | Δ4 |
| **Overall activity** | | Δ5 | Δ6/Δ8 | Δ5 | Δ6/Δ8 | Δ5 | Δ6/Δ8 | Δ5 | Δ6/Δ8 | Δ6/Δ8 | Δ5 | |

## 4 DISCUSSION

Genome and gene duplication have been recognized as decisive events for vertebrate evolution, both providing spare genetic material for adaptive selection, mutation, and genetic drift (Wagner, 1998; Lynch *et al.*, 2000; Holland, 2003). Evolutionary and functional novelties commonly arise from duplicate genes that acquire new functions – neo-functionalization, or complementary functions sub-functionalization (Glasauer *et al.*, 2014). Nevertheless, an alternative fate is more frequently reserved for duplicate genes, degeneration and ultimately loss. In fact, previous reports have found that significant gene loss occurred shortly after the 2R WGD as well as, after fish specific 3R WGD (Wagner, 1998; Lynch *et al.*, 2000; Lynch *et al.*, 2004; Blomme *et al.*, 2006; Kondrashov *et al.*, 2006; Louis, 2007; Albalat *et al.*, 2016). Although gene loss is regularly seen as less a significant player in evolution, this perspective is now shifting with recent research revealing adaptive change as consequence of gene loss (Albalat *et al.*, 2016).

### 4.1 DATA ANALYSIS REVEALS AN UNFORESEEN *FADS1* ORTHOLOGUE

Sequence and phylogenetic analysis data reveals a *Fads1* orthologue in basal jawless vertebrate Japanese lamprey, supporting the proposal that Fads1 and Fads2 originated in the vertebrate ancestor (Castro *et al.*, 2012). Also, we find that this gene orthologue is retained in basal actinopterygii, the bichir and the spotted gar that diverged prior to the teleost specific 3R WGD and in Japanese eel that diverged shortly after 3R WGD (Betancur-R. R *et al.*, 2013). However, until this date no *Fads1* orthologue has been identified in the remaining teleost species indicating that *Fads1* was most probably lost, shortly after the divergence of the elopomorha lineage (Fig. 3A). A *Fads2* orthologue was also identified in the analyzed species Japanese lamprey, bichir, spotted gar and African butterfly fish, while a Fads2 in Japanese eel had previously been identified and functionally characterized (Wang *et al.*, 2014). Importantly, in the osteoglossomorpha we find that the majority of the analyzed species present 2 *Fads2* copies including the African butterfly which are each assorted into two different clades (Fig. 1). The positioning of the clade containing the *P. buchholzi Fads2b* clade is characteristic of a direct orthology to the *Fads2* found in elopomorpha and clupeocephala lineages. On the other hand, the positioning of the *P. buchholzi Fads2a* clade is indicative these *Fads* genes are most probably retained paralogues from the 3R WGD lost in the elopomorpha and clupeocephala lineages. This hypothesis was further pursued by analyzing the genomic *locus S. formosus Fads2* genes. Here we find that the two *Fads2* copies are not tandem duplicates being placed in distinct scaffolds. However, the analysis of the neighboring genes and corresponding paralogues in *D. rerio* leaves the evolutionary origin of the genes unresolved given that the two-distinct *locus* in *S. formosus* map to two unrelated paralogous regions in *D. rerio* (supplementary material 1).

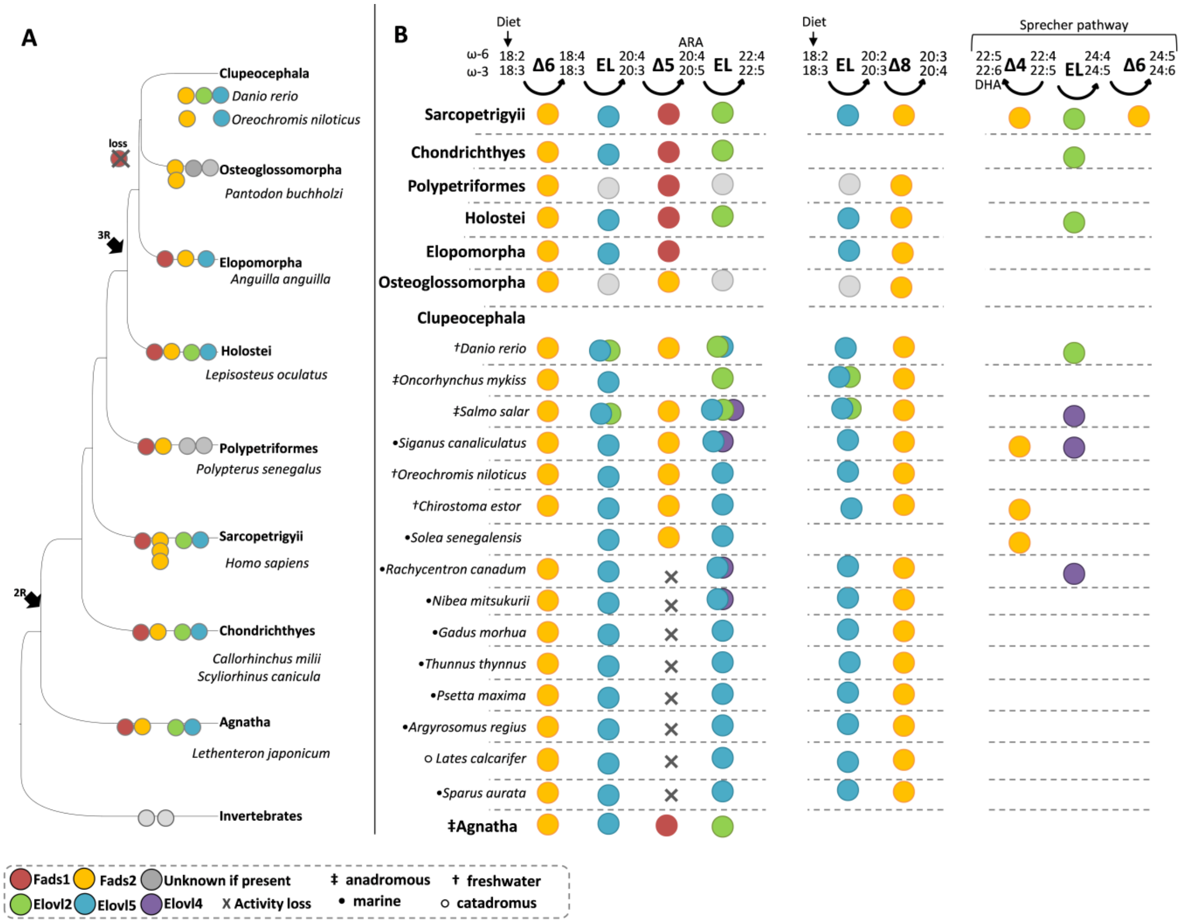


**Figure 3:** **A**- Schematic representation of the evolutionary history of the *Fads* gene family, in vertebrates. **B**- Analysis of the efficiency in completing the LC-PUFA pathway across several species. Genetic repertoire and corresponding activities of *Elovl* and *Fads* from species not characterized in the present work were retrieved from (Castro et al.; 2016).

### 4.2 LOSS OF *FADS1* PROPELS FUNCTIONAL PLASTICITY IN CLUPEOCEPHALA AND OSTEOGLOSSOMORPHA FADS2

Functional characterization reveals that the isolated desaturases present a functional phenotype similar to the one observed in sarcopterygii, presenting a Δ5 desaturase activity for Fads1 and Δ6/Δ8 desaturase activity for Fads2 (Leonard *et al.*, 2002; Castro *et al.*, 2012; Watanabe *et al.*, 2016), with the exception of *P. buchholzi* Fads2a, that presented a Δ5 desaturase activity. Interestingly, we were unable to identify a *Fads1* orthologue in *P. buchholzi* and in the genomes of the analyzed osteoglossomorpha, indicating that this gene is most probably lost in this lineage. Functional plasticity of Fads2 has been previously observed in several species of the clupeocephala lineage. For example the Δ5/Δ6/Δ8 *fads2* from *D. rerio* (Hastings *et al.*, 2001; Monroig *et al.*, 2011a), *Siganus canaliculatus* (Li *et al.*, 2010; Monroig *et al.*, 2011a), Δ5/Δ6fads2 desaturase *Oreochromis niloticus* (Tanomman *et al.*, 2013); Δ6/Δ8 *Nibea mitsukurii* (Kabeya *et al.*, 2015) yet this is the second *fads2* with solely Δ5 desaturase identified being the other in *salmo salar* (Fig. 3B) (Hastings *et al.*, 2004). In this context, it is very tempting to state that the loss of the *Fads1* orthologue in osteoglossomorpha and clupeocephala propelled a re-circuiting of the LC-PUFA biosynthesis in some species, by gene duplication and functional plasticity, to overcome the bottleneck created by the loss of Δ5 desaturase activity. *Fads2* duplication appears to be a recurrent pattern in vertebrate evolution, for example: a *Fads2* tandem duplicate *Fads3* is found most mammals (Marquardt *et al.*, 2000; Park *et al.*, 2009b; Blanchard *et al.*, 2011), in clupeocephala we also find *Fads2* duplicates, two in *C. estor* (Fonseca-Madrigal *et al.*, 2014) and *S. canaliculatus* (Li *et al.*, 2010; Monroig *et al.*, 2011a) and four *Fads2* genes in *S. salar* (Hastings *et al.*, 2004) three resulting from tandem duplication and one copy most probably from retained salmonid specific genome duplication (4R) (supplementary material 1). Gene duplication is often followed by low purifying selection, rapid sub-functionalization and or neo-functionalization (He *et al.*, 2005; Blomme *et al.*, 2006) which is seemingly the case in the referred species. However, other species that maintained one copy of *Fads2* have instead stretched its substrate preferences. For instance *D. rerio* is capable of performing Δ6/Δ5/Δ8 desaturation with single Fads2, while *C. striata* lost Δ6 desaturation activity and gained a dual Δ4/Δ5 desaturation activity (Kuah *et al.*, 2015) among various other examples (Fig 3B). These cases of functional plasticity suggest that *fads2* sequence contains fundamental elements for general desaturase activity and is permissive to be tweaked in order to retrieve other desaturation activities, while *fads1* does not appear to be permissive (Watanabe *et al.*, 2016). This has been observed in a previous study where it was possible to obtain a Fads2 with Δ5 desaturase activity and a bifunctional Fads2 Δ6/Δ5 by performing site-directed mutation. Here the replacement of a set of key residues in a Fads2 with their counterpart observed in Fads1 retrieved a Δ5 desaturase activity while the reverse experiment did not result in Fads1 with Δ6 desaturase activities (Watanabe *et al.*, 2016).

Although some species have overcome the bottleneck created by the loss of *Fads1*, by functional plasticity and/or duplication of *Fads2*, many other species remained without Δ5 desaturation activity (Fig3. B). The inability to endogenously synthesize DHA due to the lack of Δ5 desaturase activity is thought to have no significant consequence in a marine species given that these species easily obtain DHA through diet in marine rich ecosystems (Li *et al.*, 2010; Tocher, 2010). This becomes evident when observing that the majority of the Fads2 functionally characterized from marine species do not present Δ5 desaturase ability, while Fads2 from freshwater species has functionally adapted to counterbalance the nutrient poor environment (Fig 3B). In addition to environmental factors, trophic level may also drive functional plasticity, being the marine herbivore *S. canaliculatus* a remarkable example presenting two Fads2 bifunctional desaturases with Δ6/Δ5 and Δ4/Δ5 activity (Li *et al.*, 2010). Overall the dietary requirements of LC-PUFA and PUFAS vary according to a number of factors gene repertoire, and corresponding functional activities of each enzyme, environmental factors and diet.

## 5. CONCLUSION

In the present work we find a *Fads1* orthologue in Japanese lamprey indicating that *Fads1* and *Fads2* emerged in the vertebrate ancestor which is in accordance to the hypothesis previously presented findings (Castro *et al.*, 2012). Additionally, we find *Fads1* orthologues retained in basal pre 3R actinopterygii bichir and spotted gar, and in the post 3R elopomorpha Japanese ell. We find that *Fads1* was retained in post 3R teleost lineage elopomorpha and lost after the divergence of this lineage in osteoglossomorpha, ostarioclupeomorpha and euteleostei. Interestingly, we find that the majority of the osteoglossomorpha retained 2 *Fads2* genes possibly resulting from the 3R WGD, with distinct desaturase capabilities Δ5 and the second Δ6/Δ8.

## REFERENCES

Agaba, M. K., D. R. Tocher, X. Zheng, et al.; 2005; "Cloning and functional characterisation of polyunsaturated fatty acid elongases of marine and freshwater teleost fish." Comparative Biochemistry and Physiology. Part B, Biochemistry and Molecular Biology; 142;(3); 342-352.

Albalat, R. and C. Canestro; 2016; "Evolution by gene loss." Nat Rev Genet; 17;(7); 379-391.

Amores, A., J. Catchen, A. Ferrara, et al.; 2011; "Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication." Genetics; 188;(4); 799-808.

Anisimova, M., M. Gil, J.-F. Dufayard, et al.; 2011; "Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes." Syst Biol.

Betancur-R. R, Broughton RE, Wiley EO, et al.; 2013; "The Tree of Life and a New Classification of Bony Fishes. ." PLOS Currents Tree of Life.

Blanchard, H., P. Legrand and F. Pédrono; 2011; "Fatty Acid Desaturase 3 (*Fads3*) is a singular member of the Fads cluster." Biochimie; 93;(1); 87-90.

Blomme, T., K. Vandepoele, S. De Bodt, et al.; 2006; "The gain and loss of genes during 600 million years of vertebrate evolution." Genome Biology; 7;(5).

Braasch, I. and M. Schartl; 2014; "Evolution of endothelin receptors in vertebrates." General and Comparative Endocrinology; 209;(21-34).

Brunet, F. G., J.-N. Volff and M. Schartl; 2016; "Whole Genome Duplications Shaped the Receptor Tyrosine Kinase Repertoire of Jawed Vertebrates." Genome Biol Evol; 8;(5); 1600-1613.

Cañestro, C. (2012). Two Rounds of Whole-Genome Duplication: Evidence and Impact on the Evolution of Vertebrate Innovations. Polyploidy and Genome Evolution. P. S. Soltis and D. E. Soltis. Berlin, Heidelberg, Springer Berlin Heidelberg: 309-339.

Carmona-Antonanzas, G., D. R. Tocher, J. B. Taggart, et al.; 2013; "An evolutionary perspective on Elovl5 fatty acid elongase: comparison of Northern pike and duplicated paralogs from Atlantic salmon." BMC Evol Biol; 13

Castro, L. F. C., O. Gonçalves, S. Mazan, et al.; 2014; "Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history." Proceedings of the Royal Society B: Biological Sciences; 281;(1775)

Castro, L. F. C., Ó. Monroig, M. J. Leaver, et al.; 2012; "Functional Desaturase Fads1 (Δ5) and Fads2 (Δ6) Orthologues Evolved before the Origin of Jawed Vertebrates." PLoS ONE; 7;(2); e31950.

Castro, L. F. C., D. R. Tocher and O. Monroig; 2016; "Long-chain polyunsaturated fatty acid biosynthesis in chordates: Insights into the evolution of *Fads* and *Elovl* gene repertoire." Progress in Lipid Research; 62;(25-40.

Chen, S., B. H. Krinsky and M. Long; 2013; "New genes as drivers of phenotypic evolution." Nat Rev Genet; 14;(9); 645-660.

Crow, K. D., C. D. Smith, J.-F. Cheng, et al.; 2012; "An Independent Genome Duplication Inferred from Hox Paralogs in the American Paddlefish—A Representative Basal Ray-Finned Fish and Important Comparative Reference." Genome Biol Evol; 4;(9); 937-953.

Fonseca-Madrigal, J., J. C. Navarro, F. Hontoria, et al.; 2014; "Diversification of substrate specificities in teleostei Fads2: characterization of Δ4 and Δ6Δ5 desaturases of *Chirostoma estor*." Journal of Lipid Research; 55;(7); 1408-1419.

Gallardo, M. H., J. W. Bickham, R. L. Honeycutt, et al.; 1999; "Discovery of tetraploidy in a mammal." Nature; 401;(6751); 341-341.

Gallardo, M. H., G. Kausel, A. JimÉNez, et al.; 2004; "Whole-genome duplications in South American desert rodents (Octodontidae)." Biological Journal of the Linnean Society; 82;(4); 443-451.

Glasauer, S. M. K. and S. C. F. Neuhauss; 2014; "Whole-genome duplication in teleost fishes and its evolutionary consequences." Molecular Genetics and Genomics; 289;(6); 1045-1060.

González-Rovira, A., G. Mourente, X. Zheng, et al.; 2009; "Molecular and functional characterization and expression analysis of a Δ6 fatty acyl desaturase cDNA of European Sea Bass (*Dicentrarchus labrax L.*)." Aquaculture; 298;(1–2); 90-100.

Gregory, M. K. and M. J. James; 2014; "Rainbow trout (*Oncorhynchus mykiss*) *Elovl5* and *Elovl2* differ in selectivity for elongation of omega-3 docosapentaenoic acid." Biochimica et Biophysica Acta; 1841;(12); 1656-1660.

Guindon, S., J.-F. Dufayard, V. Lefort, et al.; 2010; "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." Syst Biol; 59;(3); 307-321.

Hastings, N., M. Agaba, D. R. Tocher, et al.; 2001; "A vertebrate fatty acid desaturase with Δ5 and Δ6 activities." Proc Natl Acad Sci USA; 98;(25); 14304-14309.

Hastings, N., M. K. Agaba, D. R. Tocher, et al.; 2004; "Molecular Cloning and Functional Characterization of Fatty Acyl Desaturase and Elongase cDNAs Involved in the Production of Eicosapentaenoic and Docosahexaenoic Acids from α-Linolenic Acid in Atlantic Salmon (*Salmo salar*)." Marine Biotechnology; 6;(5); 463-474.

He, X. and J. Zhang; 2005; "Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution." Genetics; 169;(2); 1157-1164.

Holland, P. W.; 2003; "More genes in vertebrates?"; J Struct Funct Genomics; 3

Jaillon, O., J.-M. Aury, F. Brunet, et al.; 2004; "Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype." Nature; 431;(7011); 946-957.

Kabeya, N., Y. Yamamoto, S. F. Cummins, et al.; 2015; "Polyunsaturated fatty acid metabolism in a marine teleost, Nibe croaker *Nibea mitsukurii*: Functional characterization of Fads2 desaturase and Elovl5 and Elovl4 elongases." Comparative Biochemistry and Physiology. Biochemistry and Molecular Biology; 188;(37-45.

Katoh, K., K.-i. Kuma, H. Toh, et al.; 2005; "MAFFT version 5: improvement in accuracy of multiple sequence alignment." Nucleic Acids Research; 33;(2); 511-518.

Katoh, K. and H. Toh; 2008; "Recent developments in the MAFFT multiple sequence alignment program." Brief Bioinform; 9;(4); 286-298.

Kondrashov, F. A. and A. S. Kondrashov; 2006; "Role of selection in fixation of gene duplications." J Theor Biol; 239

Kuah, M. K., A. Jaya-Ram and A. C. Shu-Chien; 2015; "The capacity for long-chain polyunsaturated fatty acid synthesis in a carnivorous vertebrate: Functional characterisation and nutritional regulation of a Fads2 fatty acyl desaturase with Delta4 activity and an Elovl5 elongase in striped snakehead (*Channa striata*)." Biochimica et Biophysica Acta; 1851;(3); 248-260.

Kuraku, S., A. Meyer and S. Kuratani; 2009; "Timing of Genome Duplications Relative to the Origin of the Vertebrates: Did Cyclostomes Diverge before or after?"; Molecular Biology and Evolution; 26;(1); 47-59.

Leonard, A. E., B. Kelder, E. G. Bobik, et al.; 2002; "Identification and expression of mammalian long-chain PUFA elongation enzymes." Lipids; 37;(8); 733-740.

Li, Y., O. Monroig, L. Zhang, et al.; 2010; "Vertebrate fatty acyl desaturase with Δ4 activity." Proceedings of the National Academy of Sciences; 107;(39); 16840-16845.

Lim, Z., T. Senger and P. Vrinten; 2014; "Four Amino Acid Residues Influence the Substrate Chain-Length and Regioselectivity of Siganus canaliculatus Δ4 and Δ5/6 Desaturases." Lipids; 49;(4); 357-367.

Lopes-Marques, M., R. Ozório, R. Amaral, et al.; 2017; "Molecular and functional characterization of a fads2 orthologue in the Amazonian teleost, *Arapaima gigas*." Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology; 203;(84-91.

Los, D. A. and N. Murata; 1998; "Structure and expression of fatty acid desaturases." Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism; 1394;(1); 3-15.

Louis, E. J.; 2007; "Evolutionary genetics: Making the most of redundancy." Nature; 449;(7163); 673-674.

Lynch, M. and J. S. Conery; 2000; "The Evolutionary Fate and Consequences of Duplicate Genes." Science; 290;(5494); 1151-1155.

Lynch, M. and V. Katju; 2004; "The altered evolutionary trajectories of gene duplicates." Trends in Genetics; 20;(11); 544-549.

Marquardt, A., H. Stöhr, K. White, et al.; 2000; "cDNA Cloning, Genomic Structure, and Chromosomal Localization of Three Members of the Human Fatty Acid Desaturase Family." Genomics; 66;(2); 175-183.

Moghadam, H. K., M. M. Ferguson and R. G. Danzmann; 2011; "Whole genome duplication: challenges and considerations associated with sequence orthology assignment in Salmoninae." J Fish Biol; 79;(3); 561-574.

Mohd-Yusof, N. Y., O. Monroig, A. Mohd-Adnan, et al.; 2010; "Investigation of highly unsaturated fatty acid metabolism in the Asian sea bass, Lates calcarifer." Fish Physiol Biochem; 36;(4); 827-843.

Monroig, Ó., Y. Li and D. R. Tocher; 2011a; "Delta-8 desaturation activity varies among fatty acyl desaturases of teleost fish: high activity in delta-6 desaturases of marine species." Comparative Biochemistry and Physiology. Part B, Biochemistry and Molecular Biology; 159;(4); 206-213.

Monroig, Ó., M. Lopes-Marques, J. C. Navarro, et al.; 2016; "Evolutionary functional elaboration of the Elovl2/5 gene family in chordates." Scientific Reports; 6;(20510.

Monroig, Ó., J. Rotllant, E. Sanchez, et al.; 2009; "Expression of long-chain polyunsaturated fatty acid (LC-PUFA) biosynthesis genes during zebrafish *Danio rerio* early embryogenesis." Biochimica et Biophysica Acta; 1791;(11); 1093-1101.

Monroig, Ó., D. R. Tocher, F. Hontoria, et al.; 2013; "Functional characterisation of a Fads2 fatty acyl desaturase with Δ6/Δ8 activity and an Elovl5 with C16, C18 and C20 elongase activity in the anadromous teleost meagre (*Argyrosomus regius*)." Aquaculture; 412–413;(0); 14-22.

Monroig, Ó., D. R. Tocher and J. C. Navarro (2011b). Long-chain polyunsaturated fatty acids in fish: recent advances on desaturases and elongases involved in their biosynthesis. Décimo Primer Simposio Internacional de Nutrición Acuícola. L. Cruz-Suarez, D. Ricque-Marie, M. Tapia-Salazaret al. San Nicolás de los Garza, N. L., México.

Morais, S., F. Castanheira, L. Martinez-Rubio, et al.; 2012; "Long chain polyunsaturated fatty acid synthesis in a marine vertebrate: Ontogenetic and nutritional regulation of a fatty acyl desaturase with Δ4 activity." Biochimica et Biophysica Acta; 1821;(4); 660-671.

Morais, S., O. Monroig, X. Zheng, et al.; 2009; "Highly unsaturated fatty acid synthesis in Atlantic salmon: characterization of ELOVL5- and ELOVL2-like elongases." Mar Biotechnol (NY); 11;(5); 627-639.

Morais, S., G. Mourente, A. Ortega, et al.; 2011; "Expression of fatty acyl desaturase and elongase genes, and evolution of DHA:EPA ratio during development of unfed larvae of Atlantic bluefin tuna (Thunnus thynnus L.)." Aquaculture; 313;(1–4); 129-139.

Ohno, S.; Evolution by Gene Duplication. New York, Springer Science.

Park, H. G., W. J. Park, K. S. D. Kothapalli, et al.; 2015; "The fatty acid desaturase 2 (FADS2) gene product catalyzes Δ4 desaturation to yield n-3 docosahexaenoic acid and n-6 docosapentaenoic acid in human cells." The FASEB Journal; 29;(9); 3911-3919.

Park, W. J., K. S. D. Kothapalli, P. Lawrence, et al.; 2009a; "An alternate pathway to long-chain polyunsaturates: the FADS2 gene product Δ8-desaturates 20:2n-6 and 20:3n-3." Journal of Lipid Research; 50;(6); 1195-1202.

Park, W. J., K. S. D. Kothapalli, H. T. Reardon, et al.; 2009b; "Novel Fatty Acid Desaturase 3 (FADS3) Transcripts Generated By Alternative Splicing." Gene; 446;(1); 28-34.

Pereira, S. L., A. E. Leonard and P. Mukerji; 2003; "Recent advances in the study of fatty acid desaturases from animals and lower eukaryotes." Prostaglandins, Leukotrienes and Essential Fatty Acids; 68;(2); 97-106.

Perica, M. M. and I. Delaš; 2011; "Essential Fatty Acids and Psychiatric Disorders." Nutrition in Clinical Practice; 26;(4); 409-425.

Putnam, N. H., T. Butts, D. E. K. Ferrier, et al.; 2008; "The amphioxus genome and the evolution of the chordate karyotype." Nature; 453;(7198); 1064-1071.

Robinson, L. and V. Mazurak; 2013; "N-3 Polyunsaturated Fatty Acids: Relationship to Inflammation in Healthy Adults and Adults Exhibiting Features of Metabolic Syndrome." Lipids; 48;(4); 319-332.

Santigosa, E., F. Geay, T. Tonon, et al.; 2011; "Cloning, Tissue Expression Analysis, and Functional Characterization of Two Δ6-Desaturase Variants of Sea Bass (*Dicentrarchus labrax* L.)." Marine Biotechnology; 13;(1); 22-31.

Schmitz, G. and J. Ecker; 2008; "The opposing effects of n−3 and n−6 fatty acids." Progress in Lipid Research; 47;(2); 147-155.

Seiliez, I., S. Panserat, S. Kaushik, et al.; 2001; "Cloning, tissue distribution and nutritional regulation of a Δ6-desaturase-like enzyme in rainbow trout." Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology; 130;(1); 83-93.

Session, A. M., Y. Uno, T. Kwon, et al.; 2016; "Genome evolution in the allotetraploid frog *Xenopus laevis*." Nature; 538;(7625); 336-343.

Shimeld, S. M. and P. W. H. Holland; 2000; "Vertebrate innovations." Proceedings of the National Academy of Sciences; 97;(9); 4449-4452.

Smith, J. J. and M. C. Keinath; 2015; "The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications." Genome Research; 25;(8); 1081-1090.

Tanomman, S., M. Ketudat-Cairns, A. Jangprai, et al.; 2013; "Characterization of fatty acid delta-6 desaturase gene in Nile tilapia and heterogenous expression in Saccharomyces cerevisiae." Comparative Biochemistry and Physiology. Part B, Biochemistry and Molecular Biology; 166;(2); 148-156.

Tocher, D. R.; 2003; "Metabolism and Functions of Lipids and Fatty Acids in Teleost Fish." Rev Fish Sci; 11;(2); 107-184.

Tocher, D. R.; 2010; "Fatty acid requirements in ontogeny of marine and freshwater fish." Aquaculture Research; 41;(5); 717-732.

Wagner, A.; 1998; "The fate of duplicated genes: loss or new function?"; BioEssays; 20;(10); 785-788.

Wall, R., R. P. Ross, G. F. Fitzgerald, et al.; 2010; "Fatty acids from fish: the anti-inflammatory potential of long-chain omega-3 fatty acids." Nutrition Reviews; 68;(5); 280-289.

Wang, S., Ó. Monroig, G. Tang, et al.; 2014; "Investigating long-chain polyunsaturated fatty acid biosynthesis in teleost fish: Functional characterization of fatty acyl desaturase (Fads2) and Elovl5 elongase in the catadromous species, Japanese eel *Anguilla japonica*." Aquaculture; 434;(57-65.

Watanabe, K., M. Ohno, M. Taguchi, et al.; 2016; "Identification of amino acid residues that determine the substrate specificity of mammalian membrane-bound front-end fatty acid desaturases." Journal of Lipid Research; 57;(1); 89-99.

Zheng, X., I. Seiliez, N. Hastings, et al.; 2004; "Characterization and comparison of fatty acyl Δ6 desaturase cDNAs from freshwater and marine teleost fish species." Comparative Biochemistry and Physiology. Biochemistry and Molecular Biology; 139;(2); 269-279.

## SUPPLEMENTARY MATERIAL

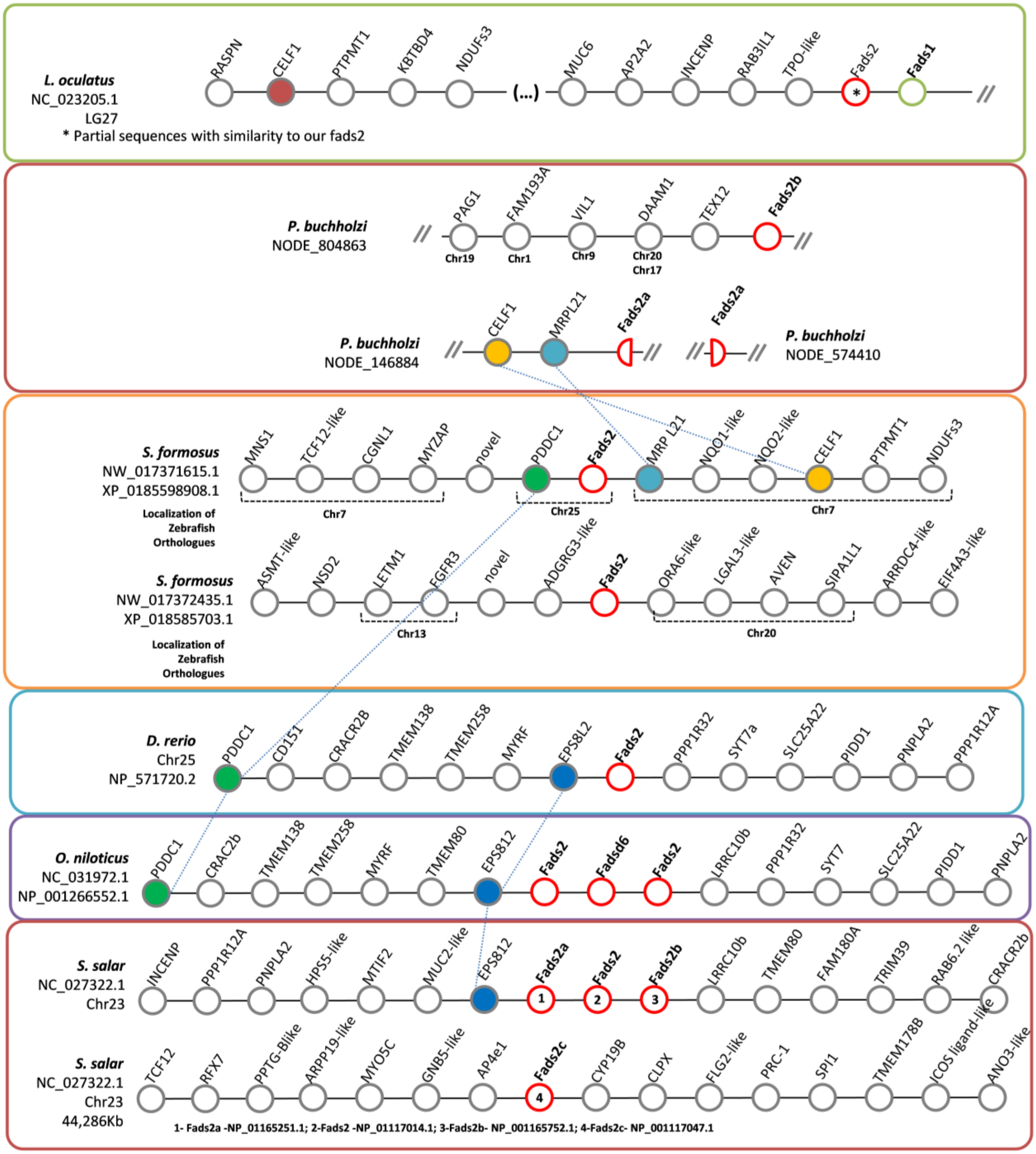


**Figure1:** Synteny maps of the *fads locus* in teleost fish

# CHAPTER V

## β-OXIDATION

# Chapter V – β-Oxidation

## V.1 The origin and diversity of CPT1 genes in vertebrate species

Mónica Lopes-Marques, Inês L. S. Delgado, Raquel Ruivo, Yan Torres, Sri B. Sainath,

Eduardo Rocha, Isabel Cunha, Miguel M. Santos, L. Filipe C. Castro

**PLoS One**

(DOI: 10.1371/journal.pone.0138447)

RESEARCH ARTICLE

# The Origin and Diversity of *Cpt1* Genes in Vertebrate Species

**Mónica Lopes-Marques[1,2], Inês L. S. Delgado[1], Raquel Ruivo[1], Yan Torres[1,2], Sri Bhashyam Sainath[1], Eduardo Rocha[1,2], Isabel Cunha[1], Miguel M. Santos[1,3], L. Filipe C. Castro[1,3]***

**1** CIIMAR, Interdisciplinary Centre of Marine and Environmental Research, CIMAR Associate Laboratory, UPorto–University of Porto, Porto, Portugal, **2** ICBAS, Abel Salazar Biomedical Sciences Institute, University of Porto, Porto, Portugal, **3** Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal

* filipe.castro@ciimar.up.pt

**Citation:** Lopes-Marques M, Delgado ILS, Ruivo R, Torres Y, Sainath SB, Rocha E, et al. (2015) The Origin and Diversity of *Cpt1* Genes in Vertebrate Species. PLoS ONE 10(9): e0138447. doi:10.1371/journal.pone.0138447

**Editor:** Hubert Vaudry, University of Rouen, France, FRANCE

**Received:** March 18, 2015

**Accepted:** August 31, 2015

**Published:** September 30, 2015

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** The authors received no specific funding for this work.

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

## Abstract

The *Carnitine palmitoyltransferase I* (*Cpt1*) gene family plays a crucial role in energy homeostasis since it is required for the occurrence of fatty acid β-oxidation in the mitochondria. The exact gene repertoire in different vertebrate lineages is variable. Presently, four genes are documented: *Cpt1a*, also known as *Cpt1a1*, *Cpt1a2*; *Cpt1b* and *Cpt1c*. The later is considered a mammalian innovation resulting from a gene duplication event in the ancestor of mammals, after the divergence of sauropsids. In contrast, *Cpt1a2* has been found exclusively in teleosts. Here, we reassess the overall evolutionary relationships of *Cpt1* genes using a combination of approaches, including the survey of the gene repertoire in basal gnathostome lineages. Through molecular phylogenetics and synteny studies, we find that *Cpt1c* is most likely a rapidly evolving orthologue of *Cpt1a2*. Thus, *Cpt1c* is present in other lineages such as cartilaginous fish, reptiles, amphibians and the coelacanth. We show that genome duplications (2R) and variable rates of sequence evolution contribute to the history of *Cpt1* genes in vertebrates. Finally, we propose that loss of *Cpt1b* is the likely cause for the unusual energy metabolism of elasmobranch.

## Introduction

Long chain fatty acids are vital players in energy homeostasis since they undergo catabolism through the β-oxidation pathway in the mitochondria. Given that the inner mitochondrial membrane is only permeable to acyl groups if linked to carnitine, fatty acid uptake requires the action of carnitine palmitoyltransferase (CPTs). This system comprises two proteins with reverse functions, CPT1 and CPT2, residing in the outer and inner mitochondrial membranes, respectively [1]. CPT1 is the rate-limiting enzyme in the trans-esterification of acyl groups from coenzyme A (CoA) to carnitine due to its sensitivity and inhibition by malonyl-CoA, an intermediate of fatty acid synthesis [2].

In mammals CPT1 enzymes are encoded by three separate genes designated *Cpt1a*, *Cpt1b* and *Cpt1c*, each expressed in different tissue compartments [1, 3–5]. *Cpt1a*, known as the liver-expressing enzyme, is found also in other tissues [6]. The second isoform, *Cpt1b*, is primarily expressed in cardiac and skeletal muscle, hence termed muscle specific, although it can also be detected in testis and adipose tissue [1, 3, 5]. A more divergent *Cpt1* gene was described and named *Cpt1c*. Commonly designated as the brain isoform, it is expressed mostly in the hypothalamus but residual levels can also appear in the ovary, testis and intestine [1, 7].

The evolution and orthology assignment of vertebrate *Cpt1* genes has posed complex questions. Orthologs of *Cpt1a*, also referred as *Cpt1a1* (see below), and *Cpt1b* have been previously identified in most vertebrate lineages [5, 8–11]. As for *Cpt1c*, the origin and function has remained difficult to elucidate. The prevailing consensus considers that *Cpt1c* is a recent gene duplicate that emerged in the mammalian lineage [2, 5, 12], probably acting as a malonyl-CoA targeted energy-sensor [2]. Recently, Ka and collaborators (2013) suggested that the sauropsid *Cpt1b* is pro-orthologous to mammalian *Cpt1b* and *Cpt1c* [8].

A further line of evolutionary complexity results from the identification of two extra *Cpt1* genes designated *Cpt1a2* alpha and beta, so far identified uniquely in teleosts [5, 9]. Their phylogenetic positioning suggests that they are a subfamily of *Cpt1a* [5, 9].

Here we re-examine the repertoire and evolutionary history of *Cpt1* genes in vertebrate species. By means of comparative genomics, phylogenetics and sampling of a basal vertebrate lineage, the chondrichthyans, we provide important insights into the evolution of the *Cpt1* gene family.

## Material and Methods

### Database identification and collection of Cpt1 genes

Using the *H. sapiens* CPT amino acid sequences, blastp and tblastn searches were performed in NCBI and Ensembl databases in order to identify and retrieve sequences from the following species: *Mus musculus*, *Sus scrofa*, *Monodelphis domestica*, *Gallus gallus*, *Falco peregrinus*, *Anolis carolinensis*, *Xenopus tropicalis*, *Latimeria chalumnae*, *Danio rerio*, *Takifugu rubripes*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Oreochromis niloticus*, *Tetraodon nigroviridis*, *Tachysurus fulvidraco*, *Callorhynchus milii*, *Drosophila melanogaster* and *Ciona instestinalis*. For *Leucoraja erinacea* Blast searches were performed on the existing genome assemblies (Build 2) and transcriptomic assemblies (Build 2) available at SkateBase [13] (S1 Table).

### Phylogenetic analysis

Sequence alignment was performed using MAFFT with the L-INS-i method [14]. The final alignment with 52 sequences was curated in BioEdit version 7.2.5 [15] with the removal of all columns containing gaps (S1 Fig), leaving an alignment with 655 gaps free sites for phylogenetic analysis. The original file with the sequence alignment containing gaps was also maintained for further phylogenetic analysis (S2 and S3A Figs). To determine the best evolutionary model of amino acid substitution, the sequence alignments were submitted to the ProtTest 2.4 server, resulting in a LG+I+G+F model [16]. Maximum Likelihood trees were reconstructed using PhyML 3.0 [17]. Branch support was assessed with aBayes [18]. Supplementary phylogenetic analysis using bayesian inference and neighbor-joining methods were conducted with the initial sequence alignment without gaps. Methods are described in S3B, S3C and S3D Fig. The resulting trees were visualized in Fig Tree V1.3.1 and rooted with the *Cpt1* homologues of *D. melanogaster* and *C. intestinalis*.

### Comparative genomics and neighbouring gene families

Comparative synteny maps were constructed with Ensembl comparative genomics pipeline, using as reference the latest available genome assemblies (Ensembl release 80—May 2015) for the following species: *H*. *sapiens* (GRCh38.p2), *M*. *domestica* (monDom5), *G*. *gallus* (Galgal4), *A*. *carolinensis* (AnoCar2.0), *X*. *tropicalis* (JGI_4.2), *L*. *chalumnae* (LatCha1) and *D*. *rerio* (GRCz10). The *F*. *peregrinus* data was collected from the latest assembly *F*. *peregrinus* v1.0 available in NCBI. For each species we analysed the genomic location of each *Cpt* gene, as well as, the five contiguous flanking genes to each side of the target gene, when possible. Following the assembly of the synteny maps, we proceeded to identify and localize the corresponding human orthologues of non-conserved neighbouring genes. Orthology was determined through the Ensembl orthologue-paralogue pipeline and our own phylogenetic analysis (not shown). Finally, synteny maps and annotated orthologues were then used to infer the localization of the ancestral *Cpt* gene in the reconstructed genome of the vertebrate ancestor using as reference the reconstruction presented by Nakatani and colleagues [19]. Synteny statistics was performed using CHSminer v1.1 [20]; input data was automatically retrieved from ensemble release 64, statistical analysis was performed for *H*. *sapiens* vs *A*. *carolinesis* and *H*. *sapiens* vs *D*. *rerio* and *H*. *sapiens* vs *X*. *tropicalis* when possible. If not indicated otherwise search parameters maintained as default maximal gap = < 30 and size> = 2. To further support synteny analysis we selected two flanking genes from each *Cpt1 locus* with representation in the majority of lineages analysed if not all, and performed phylogenetic analysis to address the orthology of the sequences (methods described in S4, S5 and S6 Figs).

## Polymerase chain reaction (PCR) and gene expression analysis

Tissues were collected from *Leucoraja erinacea* obtained from Woods Hole, USA (kind gift from Neelakanteswar Aluru). Procedures were approved by the Animal Care and Use Committee of the Woods Hole Oceanographic Institution. Total RNA was extracted using the illustra RNAspin Mini kit (GE Healthcare, UK). The RNA extraction process included an on-column DNase I treatment (provided in the kit). RNA integrity was assessed on a 1% agarose TAE gel stained with GelRed™ nucleic acid stain (Biotium, Hayward, CA, USA). The Quant-iT™ RiboGreen® RNA Assay Kit (Life Technologies, Carlsbad, CA, USA) was used to measure total RNA concentration. Reverse transcription reactions were performed with the iScript cDNA Synthesis Kit (Bio-Rad, Hercules, CA, USA). Primers targeting *Cpt1a* and *Cpt1c* genes were designed with Primer3 Software [21], using the unassembled genome sequence from *L*. *erinacea* [13]. PCR was performed with Phusion Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific, USA). Reactions were set up for a final volume of 20 μl, sense and anti-sense primer concentrations of 500 nM and 0.8 μl of template cDNA using the following general protocol: initial denaturation at 98°C for 10 seconds, a 3-step cycle including an denaturation at 98°C for 1 second, annealing for 5 seconds at a primer set specific temperature (58–61°C) and extension at 72°C during a predicted product size appropriate time (5–40 seconds) for 40 cycles and a final extension at 72°C for 1 minute. PCR products were separated by electrophoresis in 1% agarose gel. Amplification products were excised from gel and cleaned with the GRS PCR & Gel Band Purification Kit (GRiSP, Portugal) and sequenced at STABVIDA (Portugal). The resulting full ORF nucleotide sequences were deposited in GenBank: *Cpt1a* (KF570112) and *Cpt1c* (KF570111).

# Results

## Phylogenetic analysis of the Cpt1 gene family reveals three ancestral clades

We began by retrieving annotated and non-annotated CPT1-like protein sequences from genome databases of species representing all major vertebrate lineages (S1 Table). We next performed molecular phylogenetic analysis to address the overall evolutionary relationships of *Cpt1* genes. Phylogenetic analyses performed with both sequence alignments one containing gaps (S1 Fig) and the second without gaps (S2 Fig) rendered trees with similar overall topology (Fig 1 and S3A Fig). The inferred ML trees place invertebrate *Cpt1* genes outside a monophyletic group containing all vertebrate sequences (Fig 1). The later were divided into three well-supported groups encompassing *Cpt1a*, *Cpt1b*, and *Cpt1a2*/*Cpt1c* sequences respectively (Fig 1). The *Cpt1a* and *Cpt1b* clades were found to include sequences from teleosts, amphibians, coelacanth, birds and reptiles, and mammals. Contrary to previous findings, *Cpt1a2* is not unique to teleosts. Orthologues were identified in the *X*. *tropicalis*, *A*. *carolinensis*, and *L*. *chalumnae* (Fig 1). Surprisingly, the mammalian *Cpt1c* sequences robustly group with the *Cpt1a2* clade (see below). Additionally, mammalian *Cpt1c* orthologues are also apparently the least conserved, as indicated by their longer branch-lengths in the tree (Fig 1). Despite our searches, we were unable to locate an orthologue of *Cpt1c* in the available avian genomes, also confirmed by others in recent release of various avian genomes [22]. The *C*. *milii Cpt1* gene that is currently annotated as *Cpt1a* [10], robustly clusters with the *Cpt1c* group in our analysis (Fig 1).

## Synteny conservation of *Cpt1a* and *Cpt1b* loci

To further clarify the orthology/paralogy relationships of the *Cpt1* gene repertoire of the different lineages, we next examined the gene families adjacent to each *Cpt1* gene *locus* in a variety of species (Figs 2 and 3). The *Cpt1a locus* displays a high degree of synteny conservation. For example, *Mtl5* flanks *Cpt1a* in all of the examined Sarcopterygii species (Fig 2A), with the exception of *X*. *tropicalis* whose genome assembly at this *locus* is still very poor. In the paralogous *Cpt1aa* and *Cpt1ab loci* of *D*. *rerio*, the gene conservation is less evident, with the vast majority of genes having their *H*. *sapiens* orthologues mapping to chromosome 11 but at a distinct genomic region. However, adjacent to the fish *Cpt1ab* we found a novel gene family which although absent from mammals flanks the *L*. *chalumnae* and *A*. *carolinensis Cpt1a* orthologue (SIST-binding protein like) (Fig 2A). Comparative synteny statistical analysis was performed for *H*. *sapiens* vs *A*. *carolinesis* and *H*. *sapiens* vs *D*. *rerio* (Fig 2B). In both cases we find that the analysed chromosomal segments are orthologous to the corresponding *locus* in *H*. *sapiens*. In *D*. *rerio* the analysed chromosomal segment was expanded (gaps< = 100) to accommodate the highly rearranged nature of this *locus* in *D*. *rerio*. However the minimal number of genes was also proportionally increased, to maintain the statistical sensitivity. The analysis was not performed for *X*. *tropicalis* given that *Cpt1a* gene in this species is placed in an independent unplaced scaffold with no information on the neighbouring genes. Additionally, phylogenetic analysis of neighbouring genes *Mtl5* and *Sits-like*, supports the orthology of these sequences across different species (S4 Fig) and thus the common origin of this *locus*.

The gene composition of the *Cpt1b locus* also displays some degree of conservation in both Sarcopterygii and Actinopterygii species, except in *A*. *carolinensis* and *X*. *tropicalis* (Fig 3A). Even though the gene order is not exact, *Chkb*, *Arsa* and *Shank3* are all found in the proximity of *Cpt1b* in many of the examined species, providing strong support for their common origin (Fig 3A). Statistical analysis between *H*. *sapiens* vs *D*. *rerio Cpt1b locus* again indicates that these chromosomal segments are syntenic (Fig 3B). Here, statistical analysis was not performed for *A*. *carolinesis* and *X*. *tropicalis* given that *Cpt1b* gene is placed in a small scaffold in *A*. *carolinesis* or at the edge of the scaffold in *X*. *tropicalis*; in both cases lacking the minimal information regarding neighbouring genes, not allowing a confident statistical analysis of synteny. Additionally phylogenetic analysis of the neighbouring genes *Chkb* and *Arsa*, support that this genomic *locus* shares a common origin in the analysed species (S5 Fig).

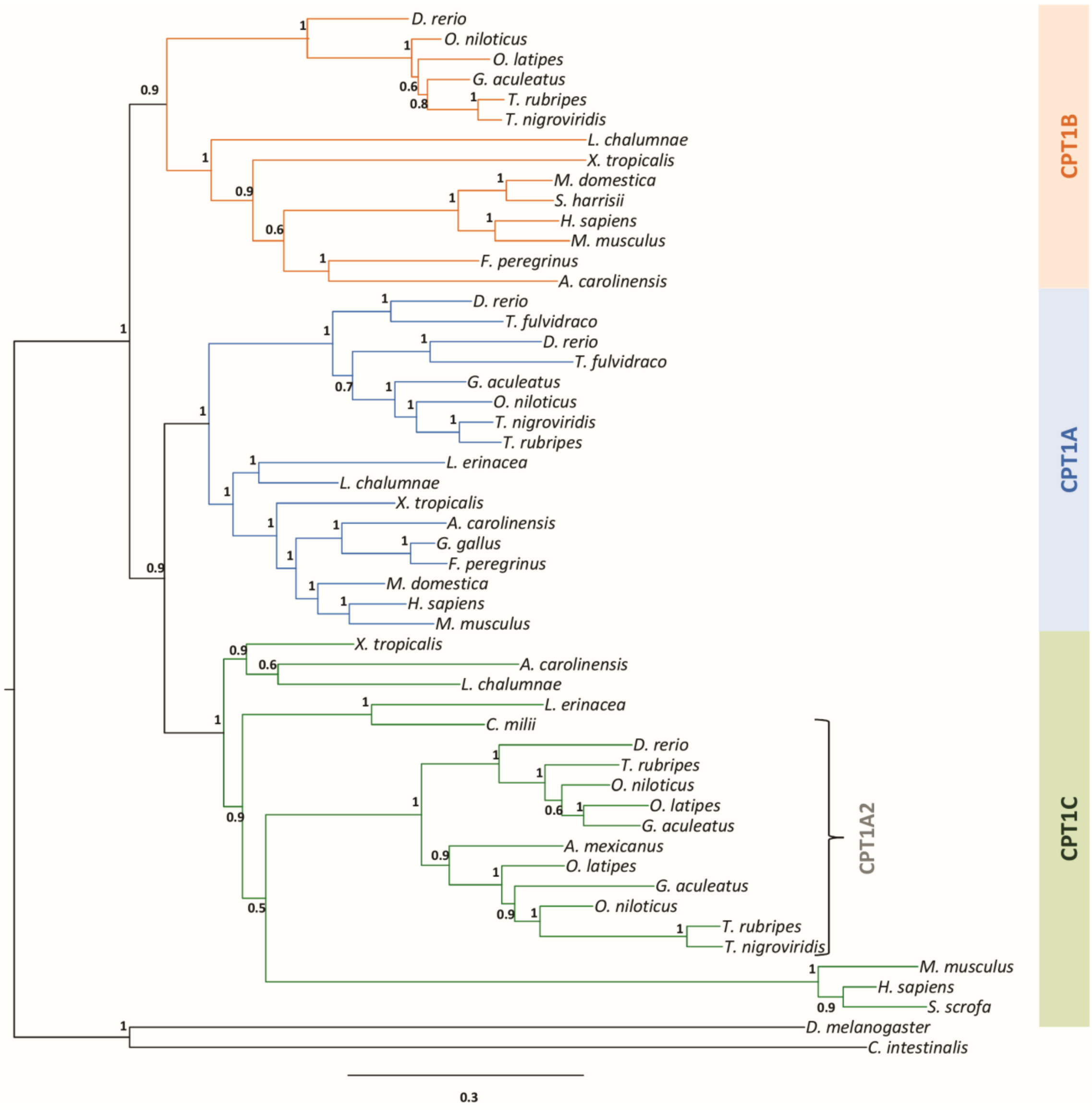
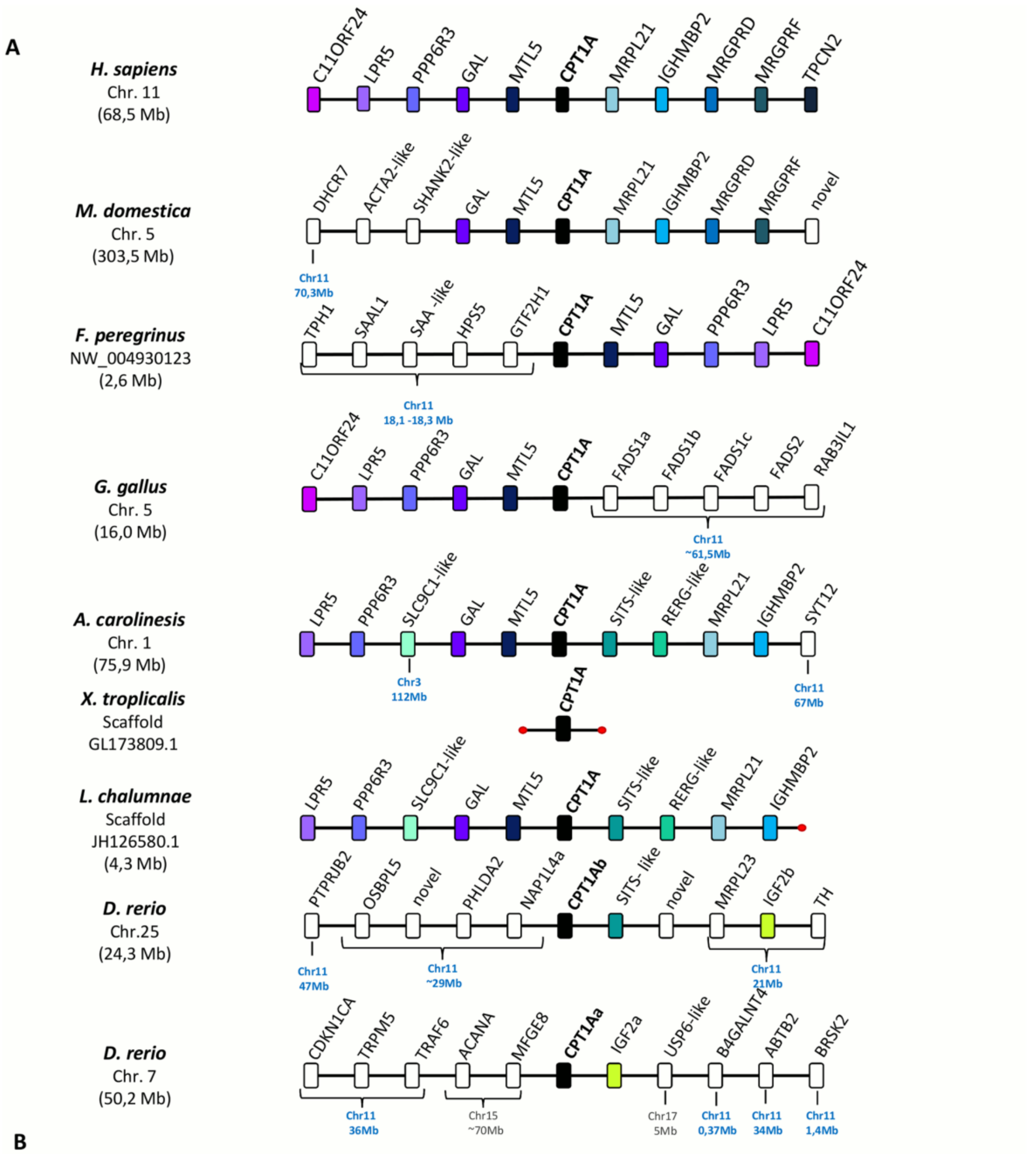


**Fig 1. Molecular phylogenetic analysis of the *Cpt1* genes by Maximum Likelihood.** Node values represent branch support using the aBayes algorithm Accession numbers for all sequences are provided in the S1 Table.

doi:10.1371/journal.pone.0138447.g001

## Locus composition supports the idea that *Cpt1a2* and *Cpt1c* genes are highly divergent orthologues

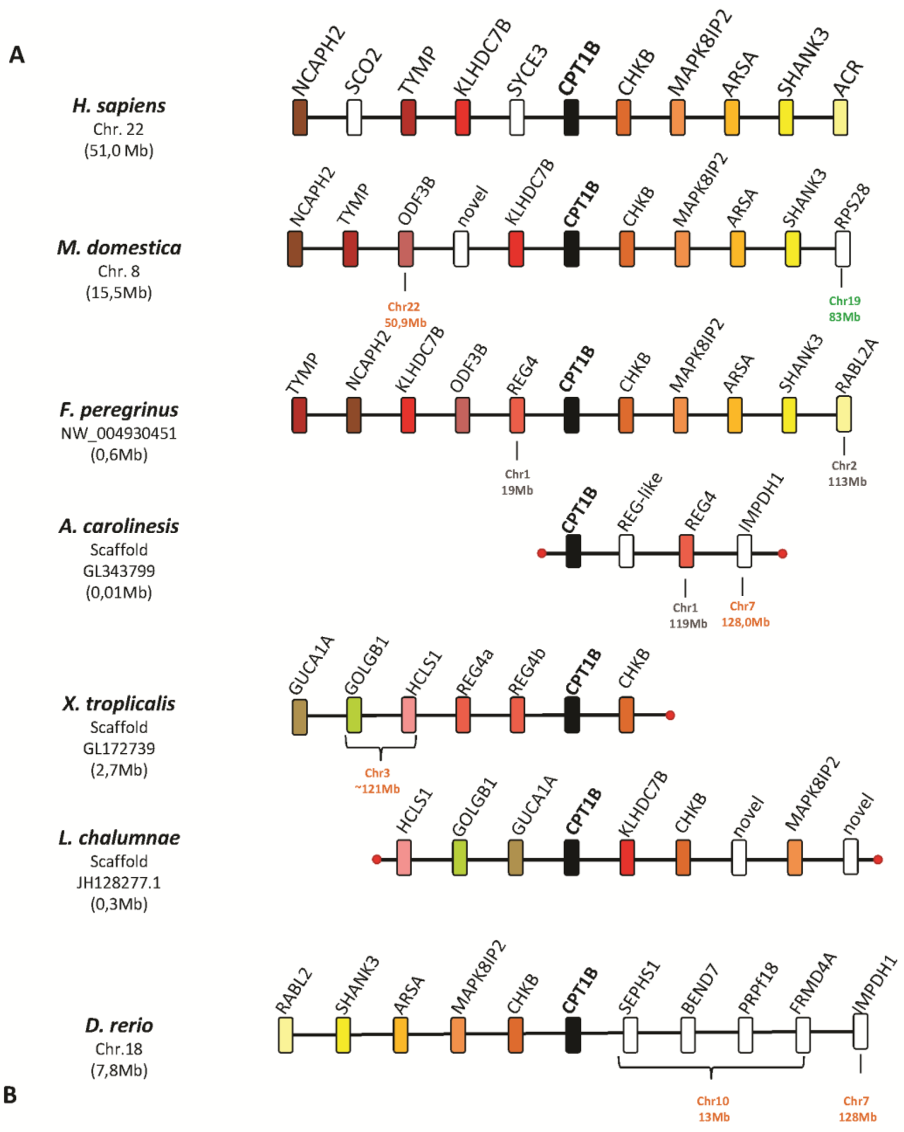
Previous reports described a new *Cpt1* gene, *Cpt1a2*, present uniquely in teleost species [5, 9] and suggested *Cpt1a2* to result from a duplication event in the teleost ancestor [9]. However,



B

| | Species | Homologous region 1 coordinates | Species | Homologous region 2 coordinates | | s | p-value |
|---|---|---|---|---|---|---|---|
| CPT1A | *A. carolinensis* | Ch 1: 75045835 - 76461782 | *H. sapiens* | Chr. 11 | 66774249 - 70035634 | 10 | 4,796E-34 |
| | *D. rerio* | Chr 25: 1186158 - 24367336 | *H. sapiens* | Chr. 11 | 57520715 - 70963623 | 18 | 7.73E-37 (100) |

**Fig 2. A. Synteny maps of *Cpt1a* gene *loci* in selected vertebrate genomes**. Chromosome (Chr.) and location in mega base pairs (Mb) is given for the gene of interest in each species. The location of the *H. sapiens* orthologue is also given for non-conserved neighbouring genes in the other species analysed Colour code denotes orthology relationships. Red dots indicate end of the chromosome or scaffold. **B. Statistical synteny analysis**. Reported p-values indicate the probability of identifying non homologous chromosomal segments, and S indicates the size of the chromosomal segment identified.



**B**

| | Species | Homologous region 1 coordinates | Species | Homologous region 2 coordinates | s | p-value |
|---|---|---|---|---|---|---|
| **CPT1B** | *D .rerio* | Chr 18: 5792077 - 7832169 | *H. sapiens* | Chr 22 50781733 - 51222091 | 10 | 2.108E-34 |

**Fig 3. A. Synteny maps of *Cpt1b* in selected vertebrate genomes**. Chromosome (Chr.) and position in mega base pairs (Mb) locations are given for the gene of interest in each species. The location of the *H. sapiens* orthologue is also given for non-conserved neighbouring genes in the other species analysed. Red dots indicate end of the chromosome or scaffold. **B. Statistical synteny analysis**. Reported p-values indicate the probability of identifying non homologous chromosomal segments, and S indicates the size of the chromosomal segment identified.

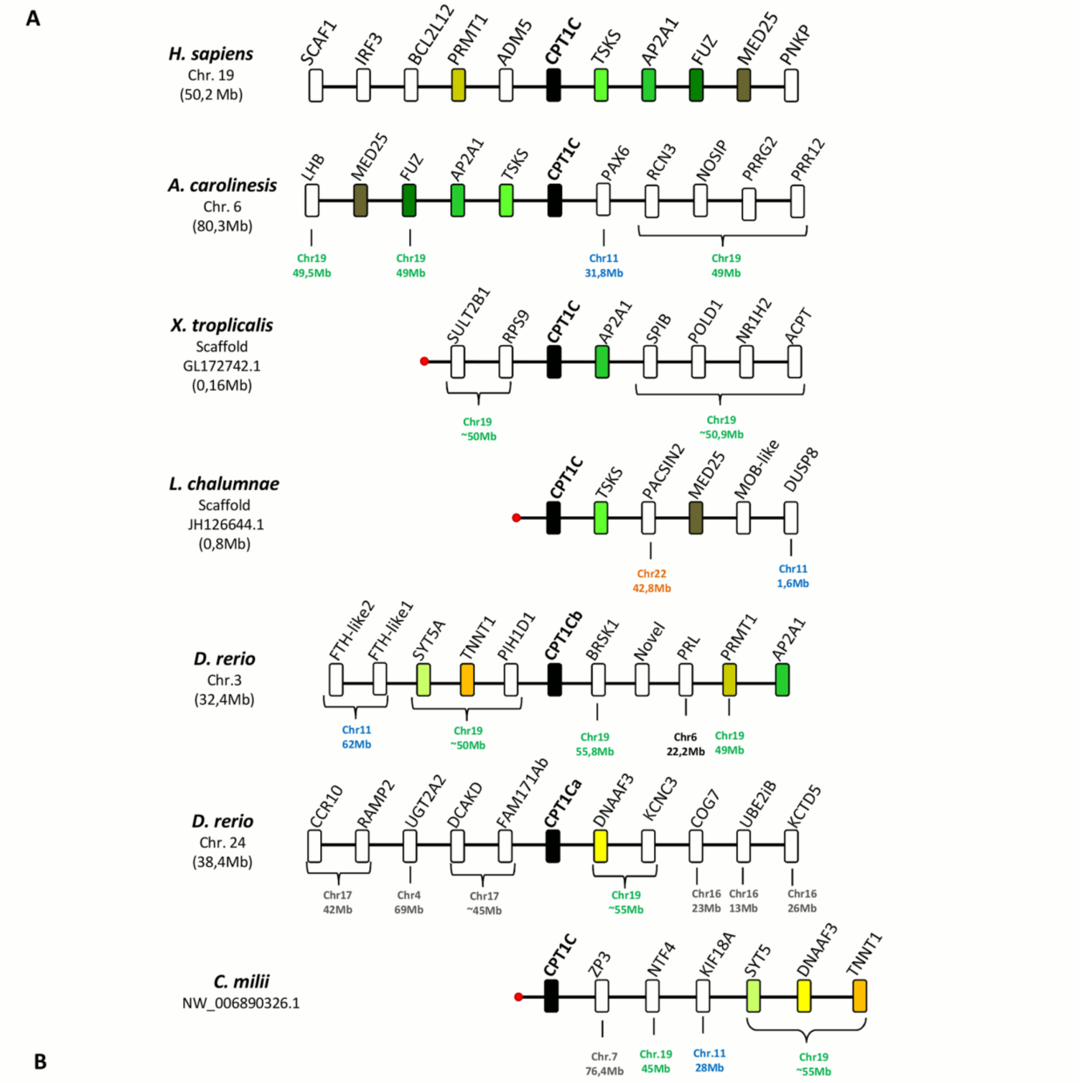
we have found orthologues, on the basis of phylogenetics, in all examined gnathostome species, except birds (Fig 1). To shed light into its evolutionary origin, we proceeded to investigate the *Cpt1a2* gene *loci* composition (Fig 4A). The *A. carolinensis* gene is flanked by *Tsks* similarly to *L. chalumnae*, while *Ap2a1* is also found close to *Cpt1a2* in all examined species, with the exception of *D. rerio Cpt1ca* and *C. milii Cpt1c* (Fig 4A). Nonetheless, we find that neighbouring genes such as *Dnaaf3*, *Kcnc3* (in *D. rerio Cpt1ca) and Ntf4* (in *C. milii*) have their human orthologues localizing to the *CPT1C locus*, establishing a conserved synteny within the analysed species (Fig 4A). In effect, detailed analysis shows that *Cpt1c* and *Cpt1a2* share a similar *locus* (Fig 4A), irrespective of the species where they occur. Additionally statistical analysis of *Cpt1c locus* synteny calculated for *H. sapiens* vs *A. carolinesis*, *X. tropicalis* and *D. rerio*, resulted in highly significant p-values in all cases (Fig 4B), indicating that these chromosomal segments are orthologous in the analysed species (Fig 4B). Both the phylogenetic and synteny analyses indicate that *Cpt1c* and *Cpt1a2* are most likely highly divergent orthologues. Thus, we propose that *Cpt1a2* from non-mammalian species should be renamed to *Cpt1c*. The occurrence of two genes in teleosts most likely results from the 3R teleost specific genome duplication [23]. In effect, *Fam171A2b* which flanks the *D. rerio Cpt1ca* has a teleost specific paralogue localizing to chromosome 3, the *locus* of origin of *Cpt1cb* (not shown).

## Cartilaginous fish have *Cpt1a* and *Cpt1c* orthologues

A *Cpt1a* gene has been recently described in a basal gnathostome, the *C. milii* [10]. However, both our phylogenetic and synteny analyses suggest that this is a *Cpt1c* gene (Figs 1 and 4). To clarify the complement of *Cpt1* genes in basal vertebrate lineages, we examined the repertoire of *Cpt1* genes in the *L. erinacea* and *C. milii*. Blast searches to the genome sequence and the transcriptome of both species identified two complete/incomplete *Cpt1*-like genes in both species. Phylogenetic analysis indicates that *L. erinacea* has *Cpt1a* and *Cpt1c* orthologues (Fig 1). These findings are also confirmed by the analysis of the *Cpt1c locus* composition in *C. milii* (Fig 4A) [10]. Careful inspection shows that the gene occurs at a similar genomic location to the *H. sapiens Cpt1c* (Fig 4A). Additionally phylogenetic analysis of neighbouring genes *Tnnt1* and *Dnaaf3* (S6 Fig) supports previous statistical analysis indicating that this chromosomal segment is orthologous between *H. sapiens* and *D. rerio* and allows us to extend this conclusion to the *C. milii Cpt1c locus*. Despite intensive searches to the genome sequence of the *C. milii*, as well as, with degenerate primer PCR in *L. erinacea*, we failed to isolate *Cpt1b* orthologues (not shown).

## Discussion

The conversion of long chain fatty acids into acylcarnitines, a fundamental step in the transport of long chain fatty acids to the mitochondria for β-oxidation, is catalyzed by CPT1. Thus, this enzyme plays an essential role in energy homeostasis, since it regulates fatty acid import for subsequent oxidation. Here, we set out to reassess the evolutionary history of *Cpt1* genes in vertebrate history, paying special attention to a basal gnathostome lineage, the chondrichthyans. These are known to have an unusual energetic metabolism without fatty acid oxidation in both skeletal and cardiac muscle [24]. Additionally, *Cpt1c* a so-called mammalian specific gene has an unclear origin and function. Several evolutionary models have been put forward to account for the reported *Cpt1* gene diversity in vertebrate lineages (Fig 5). Morash and co-workers (2010) proposed that a duplication in the ancestor of both fish and mammals gave rise to the *Cpt1a* and *Cpt1b* isoforms [9], (Fig 5 model 1), with a subsequent duplication generating *Cpt1a1* and *Cpt1a2* isoforms after the divergence of teleost fish; in an alternative scenario the *1a2* isoform was secondarily lost in mammals while retained in teleosts [9]. Extra specific genome duplications that took place in teleosts (e.g. 3R and 4R) would be responsible for the



B

| | Species | Homologous region 1 coordinates | Species | Homologous region 2 coordinates | | s | p-value |
|---|---|---|---|---|---|---|---|
| CPT1C | *A. carolinensis* | Chr 6: 78845194 - 80632227 | *H. sapiens* | Chr 19 | 49497156 - 51228979 | 29 | 7.626E-86 |
| | *X. tropicalis* | GL172742: 32955 - 1074361 | *H. sapiens* | Chr 19 | 48373723 - 51298481 | 12 | 5.89E-39 |
| | *D. rerio* | Chr 24: 38133739 - 40979232 | *H. sapiens* | Chr 19 | 48958766 - 51222707 | 10 | 1.322E-29 |
| | *D. rerio* | Chr 3: 29620402 - 32678009 | *H. sapiens* | Chr 19 | 48972289 - 55823901 | 31 | 3.28E-81 |

**Fig 4. Synteny maps of *Cpt1c* in selected vertebrate genomes.** Chromosome (Chr.) and position in mega base pairs (Mb) locations are given for the gene of interest in each species. The location of the *H. sapiens* orthologue is also given for non-conserved neighbouring genes in the other species analysed. Red dots indicate end of the chromosome or scaffold. Mapping data from the *C. milii* derived from [10]. B- **Statistical synteny analysis**. Reported p-values indicate the probability of identifying non homologous chromosomal segments, and S indicates the size of the chromosomal segment identified.

doi:10.1371/journal.pone.0138447.g004

higher number of *Cpt1* genes in fish species (e.g. *Cpt1a1a* and *Cpt1a1b*) [9]. Nevertheless, this proposal did not address the origin and evolution of the puzzling *Cpt1c* gene, nor did it provide clear insight into the duplication history of *Cpt1a1* and *Cpt1a2* genes. So far, *Cpt1c* has been largely recognized as a mammalian novelty with no orthologues identified in non-mammalian genomes [7, 25]. On the basis of phylogenetics and chromosomal mapping of the *G. gallus Cpt1b* gene, it was proposed that *Cpt1c* and *Cpt1b* emerged in mammalian ancestry from the duplication of a *Cpt1b/c* gene after the divergence of sauropsids (Fig 5 model 2) [8]. Thus, the sauropsid *Cpt1b* would be pro-orthologous of mammalian *Cpt1b* and *Cpt1c* [8].

We have put these evolutionary scenarios to the test by comprehensively mining the genomes of extant species, representative of major vertebrate lineages, for phylogenetic and synteny analyses, in particular the chondrichthyans for which no information was available. We began by undertaking molecular phylogenetics and the recovered tree topology identified three well-supported groups (*Cpt1a*, *Cpt1b*, and *Cpt1c/Cpt1a2*) in contrast to previous reports [5, 8, 9]. We also found that *Cpt1a* was present in all of the examined lineages, with its origin dating to the vertebrate ancestor. Interestingly, the gene previously designated as *Cpt1a* in the *C. milii* fails to group here, and instead is part of the *Cpt1c/1a2* group. In addition, we were able to identify a *Cpt1a* orthologue in a cartilaginous species.

We found that, in phylogenetic trees, mammalian *Cpt1c* genes branch together with the previously designated *Cpt1a2* genes, but with longer branch lengths. Thus, *Cpt1c* could be a highly divergent *Cpt1* gene without any counterpart in non-mammalian species, or a divergent orthologue of a described *Cpt1* gene. To test these possibilities we examined the synteny of *Cpt1* gene *loci*. *Cpt1a* and *Cpt1b loci* are conserved across the tested species, a clear indication that they emerged early in vertebrate evolution. Strikingly, we also found that mammalian *Cpt1c* and non-mammalian *Cpt1a2* have a similar *loci* composition, again suggesting that they are highly divergent orthologues. Our analysis allowed the simultaneous clarification of the origin of both mammalian *Cpt1c* and non-mammalian *Cpt1a2*.

*Cpt1a* and *Cpt1b* are located in genomic regions related by genome duplications in vertebrate ancestry, the so-called 2R WGD (Fig 6A) [26, 27]. Interestingly, the genomic region harbouring *Cpt1c* is part of the same linkage group [10, 26]. In this context, we put forward a model that includes the duplication of a single copy *Cpt1* gene in the ancestor of vertebrates as a result of 2R, with the retention of 3 genes and the loss of a fourth paralogue (Fig 6B). This gene complement expanded in teleosts with the lineage independent genome duplications, 3R and 4R. Based on the present data we suggest that after the divergence of sauropsids, *Cpt1c* underwent an accelerated rate of evolution and functional divergence in mammals (Fig 6B). Consequently, despite the common origin, mammalian *Cpt1c* has most likely acquired a novel function after the divergence of sauropsids. In effect, the information currently available indicates that the mammalian CPT1C function and biology is rather unique. In contrast to other CPT1 enzymes it does not localize to mitochondria but to the ER [28]. Regardless of its function, it is clear that mammalian CPT1C does not mediate mitochondrial transport of long chain fatty acids. In fact, given its oxygen demand, generation of toxic oxidative by-products and slow rate of ATP production, the brain does not rely on mitochondrial fatty acid β-oxidation, favouring glucose and liver-derived ketone bodies as source of energy [29]. Given the striking divergence of mammalian CPT1C, in both its N-terminal domain, suggested to determine protein localization and regulate activity, and C-terminal catalytic domain [30, 31], further studies are needed to elucidate their molecular function.

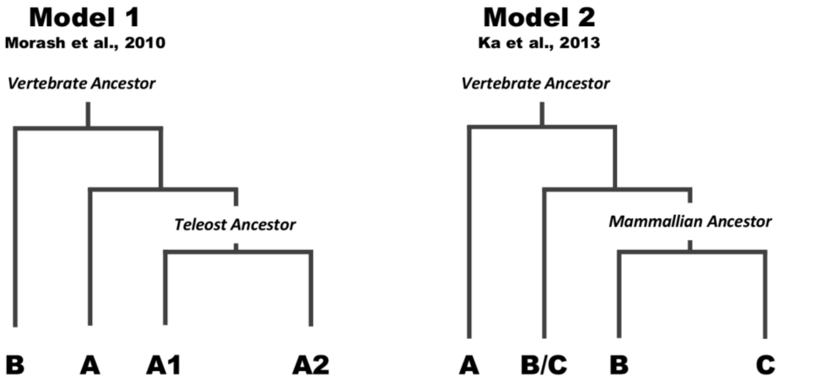


**Fig 5. Schematic representation of two evolutionary scenarios of *Cpt1* genes.** Model 1 derived from Morash et al. 2010 [9], and Model 2 derived from Ka et al. 2013 [8].

doi:10.1371/journal.pone.0138447.g005

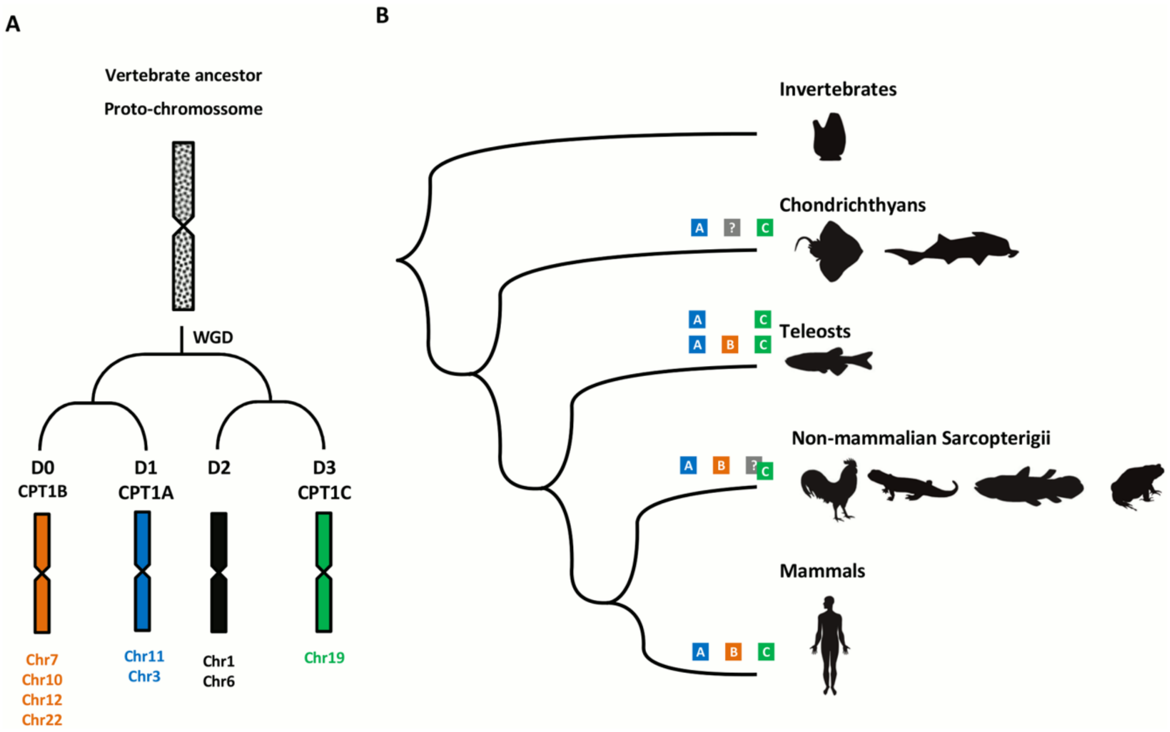


**Fig 6. Linkage group of Cpt1 genes upon genome duplications of the ancestral proto-chromosome D (details from Nakatani et al. (19) (A), and the proposed evolutionary model of the Cpt1 gene family in vertebrates (B).** WGD–whole genome duplication.

Our results also strongly suggest that *Cpt1* gene retention after 2R varied in different lineages, similar to what has been described for other gene families [26, 32, 33]. For example, we were unable to find an orthologue of *Cpt1c* in birds and *Cpt1b* in chondrichthyans. Strikingly, the absence of *Cpt1b*, the "*muscle isoform*", directly correlates with the use of ketone bodies and not fatty acids as oxidative fuels in muscle of elasmobranches [24]. If confirmed, the uncommon muscle energy metabolism elasmobranch fishes would be linked to a single event of gene loss.

## Conclusion

Our approach has provided additional clarification on the evolution of *Cpt1* genes and shows that the mammalian *Cpt1c* is probably a rapidly evolving orthologue of *Cpt1a2* in non-mammalian vertebrates. We propose that *Cpt1a*, *Cpt1b* and *Cpt1c* emerged in vertebrate ancestry as the result of genome duplications. *Cpt1c* is not a mammalian innovation (though its function probably is) since synteny and phylogenetics shows that divergent orthologues can be found in other classes. We suggest that differential loss, extra lineage-specific duplications, and an accelerated rate of sequence divergence have all modelled the history of the *Cpt1* gene family in vertebrates, with consequences in energy metabolism.

## Supporting Information

**S1 Fig. MAFFT Sequence alignment with gaps.**
(PDF)

**S2 Fig. MAFFT Sequence alignment without gaps.**
(PDF)

**S3 Fig. Alternative phylogenetic analysis supporting main phylogenetic analysis.** —Maximum likelihood phylogeny (gap alignment) using aBayes for branch support (Figure A), Maximum likelihood phylogeny with 1000 bootstraps replicates (Figure B), Bayesian Phylogenetic analysis (Figure C) and Phylogenetic analysis using Neighbor-Joining method (Figure D).
(PDF)

**S4 Fig. Supporting Phylogenetic analysis supporting *Cpt1a* comparative synteny maps.**
(PDF)

**S5 Fig. Supporting Phylogenetic analysis supporting *Cpt1b* comparative synteny maps.**
(PDF)

**S6 Fig. Supporting Phylogenetic analysis supporting *Cpt1c* comparative synteny maps.**
(PDF)

**S1 Table. List of sequences used for the molecular phylogenetic analysis and respective accession numbers (GenBank or Ensembl).**
(PDF)

## Author Contributions

Conceived and designed the experiments: MMS LFCC. Performed the experiments: ML-M ID RR YT. Analyzed the data: ML-M SBS ER IC MMS LFCC. Contributed reagents/materials/analysis tools: ML-M ID RR YT SBS ER IC MMS LFCC. Wrote the paper: ML-M RR LFCC.

## References

1. Bonnefont JP, Djouadi F, Prip-Buus C, Gobin S, Munnich A, Bastin J. Carnitine palmitoyltransferases 1 and 2: biochemical, molecular and medical aspects. Mol Aspects Med. 2004; 25(5–6):495–520. Epub 2004/09/15. PMID: 15363638
2. Wolfgang MJ, Kurama T, Dai Y, Suwa A, Asaumi M, Matsumoto S, et al. The brain-specific carnitine palmitoyltransferase-1c regulates energy homeostasis. Proc Natl Acad Sci U S A. 2006; 103 (19):7282–7. Epub 2006/05/03. PMID: 16651524
3. Esser V, Brown NF, Cowan AT, Foster DW, McGarry JD. Expression of a cDNA isolated from rat brown adipose tissue and heart identifies the product as the muscle isoform of carnitine palmitoyltransferase I (M-CPT I). M-CPT I is the predominant CPT I isoform expressed in both white (epididymal) and brown adipocytes. J Biol Chem. 1996; 271(12):6972–7. PMID: 8636126
4. van der Leij FR, Huijkman NC, Boomsma C, Kuipers JR, Bartelds B. Genomics of the human carnitine acyltransferase genes. Mol Genet Metab. 2000; 71(1–2):139–53. PMID: 11001805
5. Boukouvala E, Leaver MJ, Favre-Krey L, Theodoridou M, Krey G. Molecular characterization of a gilthead sea bream (Sparus aurata) muscle tissue cDNA for carnitine palmitoyltransferase 1B (CPT1B). Comp Biochem Physiol Biochem Mol Biol. 2010; 157(2):189–97. Epub 2010/07/06.
6. Britton CH, Schultz RA, Zhang B, Esser V, Foster DW, McGarry JD. Human liver mitochondrial carnitine palmitoyltransferase I: characterization of its cDNA and chromosomal localization and partial analysis of the gene. Proc Natl Acad Sci U S A. 1995; 92(6):1984–8. Epub 1995/03/14. PMID: 7892212
7. Price NT, van der Leij FR, Jackson VN, Corstorphine CG, Thomson R, Sorensen A, et al. A Novel Brain-Expressed Protein Related to Carnitine Palmitoyltransferase I. Genomics. 2002; 80(4):433–42. PMID: 12376098
8. Ka S, Markljung E, Ring H, Albert FW, Harun-Or-Rashid M, Wahlberg P, et al. Expression of carnitine palmitoyl-CoA transferase-1B is influenced by a cis-acting eQTL in two chicken lines selected for high and low body weight. Physiol Genomics. 2013 2013-05-01 00:00:00. 367–76 p. doi: 10.1152/physiolgenomics.00078.2012 PMID: 23512741
9. Morash AJ, Le Moine CMR, McClelland GB. Genome duplication events have led to a diversification in the CPT I gene family in fish. Am J Physiol Regul Integr Comp Physiol. 2010; 299(2):R579–89. doi: 10.1152/ajpregu.00088.2010 PMID: 20519364
10. Ravi V, Bhatia S, Gautier P, Loosli F, Tay B-H, Tay A, et al. Sequencing of Pax6 loci from the elephant shark reveals a family of Pax6 genes in vertebrate genomes, forged by ancient duplications and divergences. PLoS Genet. 2013; 9(1):e1003177. doi: 10.1371/journal.pgen.1003177 PMID: 23359656
11. Skiba-Cassy S, Collin A, Chartrin P, Medale F, Simon J, Duclos MJ, et al. Chicken liver and muscle carnitine palmitoyltransferase 1: nutritional regulation of messengers. Comp Biochem Physiol B Biochem Mol Biol. 2007; 147(2):278–87. PMID: 17337350
12. Lee J, Wolfgang MJ. Metabolomic profiling reveals a role for CPT1c in neuronal oxidative metabolism. BMC biochem. 2012; 13:23. Epub 2012/10/27. doi: 10.1186/1471-2091-13-23 PMID: 23098614
13. Wang Q, Arighi CN, King BL, Polson SW, Vincent J, Chen C, et al. Community annotation and bioinformatics workforce development in concert—Little Skate Genome Annotation Workshops and Jamborees. Database. 2012;2012.
14. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 2008; 9(4):286–98. Epub 2008/03/29. doi: 10.1093/bib/bbn013 PMID: 18372315
15. Hall T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT Nucleic Acids Symp Ser. Vol. 41 (1999), pp. 95–98.
16. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005; 21(9):2104–5. Epub 2005/01/14. PMID: 15647292
17. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59(3):307–21. Epub 2010/06/09. doi: 10.1093/sysbio/syq010 PMID: 20525638
18. Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol. 2011; 60(5):685–99. Epub 2011/05/05. doi: 10.1093/sysbio/syr041 PMID: 21540409
19. Nakatani Y, Takeda H, Kohara Y, Morishita S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res. 2007; 17(9):1254–1265. doi: 10.1101/gr.6316407 PMID: 17652425
20. Wang Z, Ding GH, Yu ZH, Liu L, Li YX (2009) CHSMiner: a GUI tool to identify chromosomal homologous segments. Algorithms Mol Biol 4: 2. doi: 10.1186/1748-7188-4-2 PMID: 19146671

**21.** Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000; 132:365–86. Epub 1999/11/05. PMID: 10547847

**22.** Lovell P, Wirthlin M, Wilhelm L, Minx P, Lazar N, Carbone L, et al. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014; 15(12):565. PMID: 25518852

**23.** Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 2004; 431 (7011):946–57. PMID: 15496914

**24.** Speers-Roesch B, Treberg JR. The unusual energy metabolism of elasmobranch fishes. Comp Biochem Physiol A Mol Integr Physiol. 2010; 155(4):417–34. doi: 10.1016/j.cbpa.2009.09.031 PMID: 19822221

**25.** Wolfgang MJ, Cha SH, Millington DS, Cline G, Shulman GI, Suwa A, et al. Brain-specific carnitine palmitoyltransferase-1C: role in CNS fatty acid metabolism, food intake and body weight. J Neurochem. 2008; 105(4):1550–9. doi: 10.1111/j.1471-4159.2008.05255.x PMID: 18248603

**26.** Feiner N, Meyer A, Kuraku S. Evolution of the vertebrate Pax4/6 class of genes with focus on its novel member, the Pax10 gene. Genome Biol Evol. 2014; 6(7):1635–51. Epub 2014/06/22. doi: 10.1093/gbe/evu135 PMID: 24951566

**27.** Lundin L-G, Larhammar D, Hallböök F. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. In: Meyer A, Van de Peer Y, editors. Genome Evolution: Springer Netherlands; 2003. p. 53–63.

**28.** Sierra AY, Gratacos E, Carrasco P, Clotet J, Urena J, Serra D, et al. CPT1c is localized in endoplasmic reticulum of neurons and has carnitine palmitoyltransferase activity. J Biol Chem. 2008; 283(11):6878–85. doi: 10.1074/jbc.M707965200 PMID: 18192268

**29.** Schonfeld P, Reiser G. Why does brain metabolism not favor burning of fatty acids to provide energy? Reflections on disadvantages of the use of free fatty acids as fuel for brain. J Cereb Blood Flow Metab. 2013; 33(10):1493–9. doi: 10.1038/jcbfm.2013.128 PMID: 23921897

**30.** Cohen I, Guillerault F, Girard J, Prip-Buus C. The N-terminal domain of rat liver carnitine palmitoyltransferase 1 contains an internal mitochondrial import signal and residues essential for folding of its C-terminal catalytic domain. J Biol Chem. 2001; 276(7):5403–11. Epub 2000/11/23. PMID: 11087756

**31.** Samanta S, Situ AJ, Ulmer TS. Structural characterization of the regulatory domain of brain carnitine palmitoyltransferase 1. Biopolymers. 2014; 101(4):398–405. doi: 10.1002/bip.22396 PMID: 24037959

**32.** Castro LF, Lopes-Marques M, Wilson JM, Rocha E, Reis-Henriques MA, Santos MM, et al. A novel Acetyl-CoA synthetase short-chain subfamily member 1 (Acss1) gene indicates a dynamic history of paralogue retention and loss in vertebrates. Gene. 2012; 497(2):249–55. Epub 2012/02/09. doi: 10.1016/j.gene.2012.01.013 PMID: 22313524

**33.** Mulley JF, Holland PWH. Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. Mol Biol Evol. 2010; 27(10):2386–91. doi: 10.1093/molbev/msq121 PMID: 20463047

# SUPPLEMENTARY MATERIAL

**Supporting information 1 Fig:** MAFFT Sequence alignment with gaps

```
Hsa1C        MAEAHQAVGF RPSLTSDGAE VELSAPVLQE IYLSGLRSWK RHLSRFWNDF LTGVFPASPL SWLFLFSAIQ L-AWFLQLDP SLGLMEKIKE LLPDW-----  94
SSc1C        .......... ..P.....G. ....P.L... .......... .......... .I........ .......... .-.C...... .......... .....-----  94
Mmu1C        .......SSL LS..S..... ....S..W.. ...CA..... ...W.V.... .A..V..T.. .......T.. .-.CL..... .......... .....-----  94
Hsa1A        ........A. QFTV.P..ID LR..HEA.RQ ......H... KKFI..K.GI I...Y....S ...IVVVGVM T-TMYAKI.. ...IIA..NR T.ETA-----  94
Mmu1A        ........A. QFTV.P..ID LR..HEA.KQ .C....H... KKFI..K.GI I........S ...IVVVGVI S-SMHTKV.. ...MIA..NR T.DTT-----  94
Mdo1A        ........A. QFTV.P..ID LR..HEA.KQ ......H... KKFI..K.GI I...Y....S ...IVVVGVM S-TMYAKV.. ...IIA..NQ T.DMT-----  94
Gga1A        ........A. QFTV.P..ID LRM.HEA.KQ .....VH... KKFI..K.GI I...Y....S ...IVVVGVM S-TMYAKI.. ...IIA..NR T.DTT-----  94
Fpe1A        ........A. QFTV.P..ID LRM.HEA.KQ .....VH... KKFI..K.GI I...Y....S ...IVVVGVM S-TMYAKI.. ...IIA..NR T.DTT-----  94
Aca1A        ........A. QFTV.P..ID LRM.HEA.RQ ......Y... KKFI..K.GI I...Y....S ...IVVVGVM S-TMYAKI.. ....IA..NR T.DIT-----  94
Xtr1A        ........A. QFTV.P..ID LR..HEA.RQ ......H... KKFI..K.GI ......S..S ...IVVVGVT SASMYTKV.. .F.II...NN A.SAT-----  95
Lch1A        ........A. QFTV.P..ID LR..HEA.RQ ......H... KKFI..K.GI I...Y....S ....VVVGVV S-TMYTRV.. ...MIAR.SQ R..VT-----  94
Ler1A        ........A. QFTV.P..ID LQ..HEA.KQ V......... KRFR.YK.GI ....Y....S ....VVV.VI A-TMYARV.. ...MI...RC H..TN-----  94
Xtr1C        ........A. QFTV.PE.ID LR..HEA.KQ V......... KKFA.VK.S. I...Y....S ....VVI..M G-TLYARV.. .M.MI..... H..AS-----  94
Ler1C        ........A. QFTV.PE.ID LQ..RHA.K. .....I.... KKCN.IK.N. M...Y....S ..I.VVLTVV G-TMYTRV.. .M..IAMLR. HI.LR-----  94
Cmi1C        ........A. QFTV.PE.ID LQ..HQA.K. .....V.A.. KRCA.IR.NI V...Y....S ....VALTVV G-TMYTRV.. .M.MIG.... H..VK-----  94
Aca1C        ........A. QFTVSPE.ID LQ..HVAFK. .....V.... KKCNQLR.G. V...Y....S ....MVT..L V-TQYSR... .M.MID.... H..VS-----  94
Lch1C        ........A. QFTV.PE.ID LK..HEA.KQ VF...Q.... KR.A.LK.S. I..MY....S .....VI..M A-TLYARV.. .M.MID.... H..VKVQEFY  99
Tni1A        ........A. QFTV.P..ID LQ.CHEA.RQ V.....H... KRFI..K.GV M...Y.G..A GFMIVVGSYM SYNKYNR... .M..VV.LGQ YI.IG-----  95
Tru1A        ........A. QFTVSP..ID LQ.CHEA.RQ ......H... KRFI..K.GV M...Y.G..G GFMIVVGSYM SYNKYN.... .M.FVV.LGQ YI.IS-----  95
Oni1A        ........A. QFTV.PE.ID LH..HEA.RQ V....IH... KRFV..K.GI M...Y.G..A GFSIVVGSYM AYNK.K.... ....FT.LGQ HI.IS-----  95
Gac1A        ........A. QFTVSP..ID LQ..HEA.RQ ......H... KRFI..K.GV M...Y.G..G GFMVVVVSYM SYNKYQ.... ....IA.LGQ HM.IS-----  95
Dre1Aa       ........A. QFTV.P..ID LH.CHEA.RQ .....IH... KRFI..K.GV M...Y.G..S GLVVVLVGYM SSTKYAKI.. ...ILT.LST H..VS-----  95
Tfu1A2       ........A. QFTV.P..ID LQ.CHEA.RQ ....A.H... KRFI..K.GV M...Y.G..T GL.GVVGGYL IFTKYANI.. T..VVA.LAP H..VC-----  95
Dre1Ab       ........A. QFTVGP..ID LQ..HEA.RQ V....IH... KKFI..K.GI .N..Y.G.AP GFVLVLAGYL GR.QY.KV.. ....LF.LGN YV.IS-----  95
Tfu1A1       ........A. QFTVSP..ID LQ..HEA.RQ V.....H... KKFI..K.GI M...Y.G.AP GLMLVLAGYM GR.KYA.V.. ....VFR.GK YV.IS-----  95
Ola1Ca       ........A. QFTV.PE.ID LR..HQA.S. .....V.... KRII.IK.SV I...Y....S ....VVI..L A-TMYTRS.. .M.IIA..Q. H..VS-----  94
Oni1Cb       ........A. QFTV.P..ID LQ..HQA.T. .....M.... KRIV.VK.SV I...Y....S ....VVI..L A-TMYTRS.. .M..IA..Q. H..VS-----  94
Ame1C        ........A. QFTI.PE.ID LQ..HEA.RQ V......... KRIV.LK.NV I...Y....S ....VVI..L A-TMYTRS.. .M..IA..Q. H..AS-----  94
Gac1C        ........A. QFTVSP..ID LH..HQA.T. .....V.... KRVIGLK.SV V...Y...LS ....VVI..L A-TMYTRS.. .M..IA..Q. H..RP-----  94
Tru1C        .......... EFTVIPE..D AQ..HQA.T. ....AVN... KCII.IK.SV IR..Y....F ....VVI..L A-TMYTRS.. .M..IA..Q. N..VS-----  94
Tni1C        .......... EFTVIPE.TD AQ..HQA.T. ....AVH... KCII.IK.SV IK..Y....F ....VVI..L A-TMYTRS.. .M..IA..Q. N..VS-----  94
Ola1Cb       ........A. QFTI.PE.ID LQ..YQA.NQ .....V.... KRV..MR.RV IK..Y....S ....VAIG.L A-TIYM.S.. .M..IAE.QQ R..LS-----  94
Oni1Ca       ........A. QFTI.PE.ID LQ..YQA.NQ .....V.... KRV..VR.SL IK..Y....S ....VAIG.L A-TMYM.S.. .M..IT..QQ H..LS-----  94
Gac1C        ........A. QFTI.PE.ID LQ..HQA.NQ .F...V.... KRV..MR.RV IK..Y....S ....VVI..L A-TMYMRS.. .M..IKM.QQ H..LR-----  94
Tri1C        ........A. QFTI.PE.ID LQ..YQA.NQ .....V.... KRV..VR.SL IK..Y....S ....VSI..L A-TMYMRS.. .M..IT..QH H..LS-----  94
Dre1C        ........A. QFTISPE.IN LH..YQA.NQ .......... KRI..IK.RI IK.AY....S ....IVI..L A-TMYM.S.. .M..IA..Q. H..LS-----  94
Dre1B        ........A. QFTV.P..ID LQ..RE..KH ..I..VT... KRAI..K.SV ....Y....S ....VVI..M S-TMYARI.. .M.TID...T S..VS-----  94
Oni1B        .......... QFTVRP..VD FK..QE.IKN .....VTA.. KKAIQ.K.GV .A..Y....S ...IVVI.MM S-SLYIHI.. ...M.DA... N..YR-----  94
Gac1B        .......... QFTVRP..VD LK..QE.IKT .....VTA.. KRAIQ.K.GV .A..Y....S ...IVVI.MM S-SLYIRI.. ...MI.ALQ. N..HR-----  94
Ola1B        .......... QFTVRP..V. LK..QE.IKN ......TV.. KKAIQ.K.SV ....Y....S ...IVVI.MM S-SLYTNT.L ...MIDA... N..HR-----  94
Tru1B        .......... QFTVRP..V. LK..QE.IKN ......TA.. KRAIQ.K.GV .A..Y....S ...IVVI.MM S-SLYTRV.. ...MI.AM.. N..YR-----  94
Tni1B        .......... QFTVRP..V. LK..QE.IKN ......TA.R KRAIQ.K.GV .A..Y....S ...IVVI.MM S-SLYI.V.L ...MI.AM.. N..YR-----  94
Hsa1B        ........A. QFTV.P..VD FR..REA.KH V....IN... KR.I.IK.GI .R..Y.G..T ...VVIM.TV G-SS.CNV.I ....VSC.QR C..QGCG---  96
Mmu1B        ........A. QFTV.P..VD FR..REA.RH .....IN... KR.I.IK.GI .R..Y.G..T ...VVVM.TV G-SNYCKV.I .M..VDC.QR C..ERYG---  96
Mdo1B        ........A. QFTV.P..VD FQ..REA.KH V....IN... KR.I.IK.GI .R..Y.G..T ...VVVMTTM G-SSYCNV.L .M.MICC.RK YI.EGCC---  96
Sha1B        ........A. QFTV.P..VD FQ..REA.KH V....IN... KRFI.IK.GI .R..Y.G..T ...VVVM.TM G-SSYCNV.I .M.MICH.RK YI.EGCS---  96
Aca1B        ........A. QFTV.PE.VA FQ..REA.KQ .....VS... KR.A.LK.SI ......GT.S G..AVVAVTV G-.TYFGI.V .V..IFR.QK R..NSCR---  96
Fpe1B        ........A. QFTV.PE.LD FH..REAVRQ L..A.IS... KR.V.AK.S. ....Y....S ..MVVVM.TA G-SFYC.V.. ...MIAR.RH C..ES-----  94
Xtr1B        ........A. QFTV.P..ID LQ..HEA.RH LW....VACR KR.ITLKSSV .S..Y....S T.FAVVAMTL G-SLYGK... ....IT..NS I..GK-----  94
LchB         ---------- ---------- ---------- -----.MN.L H--VFVQ.SL ....Y....S ...VVVV.TI G-TRYVKM.. ...IIDY.RS AI.RS-----  57
Dm           .....A..A. SFAI.HE.FD INYDHE..NL VWN..V.... KR.A.AR.GV RN..Y..HIQ .LWLISAIAL G-LH.AGYQA PFN.TNR.LV H..SN-----  94
ci           .....A..A. SF.V.QE.LN .QI.HEAM.A V.F..V...R KR.T..R.RV .S....VKYT .LVG.TAIVL .-.S..SY.I TW.FKNNRTN II.PR-----  94


Hsa1C        ---GG----- -QHHGLRGVL AAALFASCLW GALIFTLHVA LRLLLSYHGW LLEPH---GA MSSPTKTWL- -ALVRIF--- ------SG-R H-PMLFSYQR 169
SSc1C        ---.A----- -...R...I. .......... .......... .......... ...T.---.V .........- -......--- ------..-. .-........ 169
Mmu1C        ---..----- -...Q.Q.F. S..V...... .......... ......H... .....---.. .........- -......--- ------..-. .-.R...F.. 169
Hsa1A        ---NC----- -MSSQTKN.V SGV..GTG.. V...V.MRYS .KV....... MFTE.---.K ..RA..I.M- -GM.K..--- ------..-. K-...Y.F.T 169
Mmu1A        ---.R----- -MSSQTKNIV SGV..GTG.. V.I.M.MRYS .KV....... MFAE.---.K ..RS.RI.M- -.M.KV.--- ------..-. K-...Y.F.T 169
Mdo1A        ---.S----- -MSSQTKNIV SGI..GTG.. VT..VAMRYS .KM....... MFAEY---.K L.RG.RI.M- -GM.K..--- ------..-. K-...Y.F.T 169
Gga1A        ---.Y----- -MSNQTQNIV SGI..GTG.. V...V.MRYS .KM....... MFAE.---.K L.AG..L.M- -T..KL.--- ------..-. K-...Y.F.T 169
Fpe1A        ---.Y----- -MSNQTQNIV SGV..GTG.. V...V.MRYS .KM....... MFAE.---.K L.AG..F.M- -...KL.--- ------..-. K-...Y.F.T 169
Aca1A        ---.Y----- -MSNQTQNIV SGV..GTG.. V....SMRYT .KM....... MFAE.---.K L.TG..I.M- -T..KL.--- ------..-. K-...Y.F.T 169
Xtr1A        ---.Y----- -MTPQTQNIV SGV..GTG.. VS..A.MRYS .KK....... MF.E.---.K L.AS.RI.M- -GM.KLL--- ------..-. K-...Y.F.T 170
Lch1A        ---TH----- -LSSQSQNIV SGV..GTG.. V..F..MRYS .KM....... MFIE.---.K VPFG.RI.I- -R..KL.--- ------.V-C K-...Y.F.T 169
Ler1A        ---.Y----- -LSNQSESI. SGL..STG.. I..V..MRQT .K........ MFV..---.K TPTSV.L.M- -LI.K..--- ------..-. K-.LTY.F.T 169
Xtr1C        ---SY----- -LS.QGQSIV S.L..STG.. F...M.MRFI .KQ....... MY.Q.---.K ..AT..L..- -...K..--- ------..-. N-...Y...A 169
Ler1C        ---RY----- -MS.HGQSMV S.M...TG.. L....VMRQI .KT....... MF.E.---.K V..T..V.F- -T..K..--- ------..-. K-...Y.F.N 169
Cmi1C        ---RC----- -LS.QGQNIM S.L...TV.. L...YMMRQI .KT....... MF.E.---.K A.NM.RI.F- -TM.K..--- ------..-. K-...Y.F.A 169
Aca1C        ---.Y----- -LSDQGMNI. S..T.STV.. L...M.MRSI .KM..C.... MY.E.---.K ..NT..I..- -...KM.--- ------A.-. K-...Y...A 169
Lch1C        QTARTKTNPT QITDKRINL. CTN.YQCGIY SEILL----- ---------- MF.Q.---NR V.LT..I.I- -T..K..--- ------..-. K-...Y...A 168
Tni1A        ---RY----- -LSTDTQKIV GGV.VGTS.. VTI.MIMRNV .KS...W... MH.R.---.S V.WTSRA.M- -L..KV.--- ------..-. K-...Y.F.N 170
Tru1A        ---RY----- -LSTDTQRIV GGV.VGTG.. VTI.MIMRNV .KS...W... MY.R.---.S I.WTSRA.M- -L..KV.--- ------..-. K-...Y.F.N 170
Oni1A        ---RY----- -MSTDSQKIV GGI.VGTS.. VTI.MIMRTV .KS...W... MQ.R.---.S L.WSSRI.M- -V..KV.--- ------..-. K-...Y.F.N 170
Gac1A        ---RY----- -MSTDSQKIV GGV.VGTG.. VSI.MIMRSV .KS...W... MYTS.---.S VAWS.RL.M- -V..KV.--- ------..-. K-...Y.F.N 170
Dre1Aa       ---KY----- -ITEDGQRIV GGV.VGTG.. I.VT.VMRN. .KY...W... MFNQ.---.T L.LK..I..- -V..KL.--- ------..-P K-...Y.F.S 170
Tfu1A2       ---KY----- -ITEDTQQIV GGI.VGTGV. T.V..FMRNM .KC...W... MYNR.---.T ..IR..I..- -...KL.--- ------..-P K-...Y.F.S 170
Dre1Ab       ---RY----- -MSEQKQMLV GGVMVGTG.. I.I..GMRTV .KG...W... MFAS.---.R .TWKIRL..- -VF.KV.--- ------..MQ T-...Y.F.N 171
Tfu1A1       ---KY----- -MSLDNQSRV GGI.VGTG.. VVI...MRSI .KG...W... MR.R.---.S LTWK.QI..- -V..KV.--- ------..-M KT.N.Y.F.T 171
Ola1Ca       ---QS----- -MSSQCQA.V S.V..STM.. LL....MRLC .KQ.....R. MF.Q.---.K ..TT..I.V- -......--- ------..-. K-.L.Y...G 169
Oni1Cb       ---QS----- -MSTQCQA.V S.V..STM.. LM....MRLC .KQ.....R. MF.Q.---.K ..TT..I.V- -......--- ------..-. K-.L.Y...G 169
Ame1C        ---.S----- -MSMQCQT.V S.V..STL.. LS....MKLC .KQ.....R. MF.Q.---.N V.TT..V.V- -M.....--- ------..-. Q-.L.Y...G 169
Gac1C        ---------- ------YVM. S.S.AVTM.. LL....MRMC .KQ.....R. MF.K.---.K ..NT..V.V- -VS.ASHRSL FVTVNQPS-. K-.L.Y...G 171
```

```
Tru1C        ---QS----- -LSTQCKVL. S.VI.STM.. LL....MRLC .KQ.....R. MF...---.K ..TT..V.V- -......--- ------..-. K-.R.Y...A  169
Tni1C        ---QS----- -LSTQCKVL. S.VI.STM.. LL....MRLC .KQ.....R. MF...---.K ..TT..V.V- -......--- ------..-. K-.R.Y...G  169
Ola1Cb       ---LHVS--- -LSAQGQTMV S.LV.STL.. LS..LA.RFC .K......Q. MF.Q.---.R V.NT..V.V- -T.L.LL--- ------..-. K-.L.Y...T  171
Oni1Ca       ---FF----- -LFTQGQTMV S.LV.STL.. LS..LA.RFC .K......Q. MF.K.---.R I.NT..V.V- -P...LL--- ------.S-. K-.L.Y...S  169
Gac1C        ---QPF---- -LSTQGQTM. S.LV.STL.. LS..LA.RFC .K......Q. MF.Q.---.R V.NT..V.V- -T.L.LL--- ------.S-. K-.L.Y...T  170
Tri1C        ---LHMS--- -LSAQGQTM. S.LV.STL.. LS..LA.RFC .K......R. MF.Q.---.R V.NF..V.V- -T.LGLL--- ------.S-. K-.L.Y...T  171
Dre1C        ---LS----- -LSPQGQTM. S.L..STL.. MS..L..RFC .K......R. MF.Q.---.H ..TK..V.A- -T..KLL--- ------.S-. K-.L.Y...T  169
Dre1B        ---EF----- -MTVQTQT.. S.I...TG.. LSV..L.RYL .KA.....A. IF.S.---.K ..YS..V..- -S..KLL--- ------..-. R-.L.Y.F.G  169
Oni1B        ---DC----- -MSVQT.A.. S.I...TG.. LF..YL.RYT .KA....... IF.S.---.K ..TS..V..- -C..KM.--- ------..-. R-.L.Y.F.A  169
Gac1B        ---DY----- -LSVQS.A.. S.I...TG.. LF..YL.RYM .KA....... IF.S.---.K ..RS..V..- -S..KM.--- ------..-. R-.L.Y.F.A  169
Ola1B        ---.Y----- -MSAQT.A.. S.I..GTG.. LF..YL.RYT .KA....... IF.S.---.K ...S..L..- -Y..KM.--- ------..-. R-.L.Y.F.A  169
Tru1B        ---DC----- -MSVQT.A.. S.I...TG.. LF..YL.RYT .KA....... IF.S.---.K ..TS..V..- -T..KM.--- ------..-. R-.L.Y.F.A  169
Tni1B        ---D.----- -MSVQT.A.V S.I...TG.. LF..YL.RYT .KA....... IF.S.---.K ..TS..V..L QT..KM.--- ------..-. R-.L.Y.F.A  171
Hsa1B        ---PY----- -.TPQT.AL. SM.I.STGV. VTG..FFRQT .K...C.... MF.M.---.K T.NL.RI.A- -MCI.LL--- ------.S-. .-...Y.F.T  171
Mmu1B        ---HF----- -GTPQTEAL. SMVI.STGV. ATG..FFRQT .K........ MF.M.---SK T.HA..I.A- -IC..LL--- ------.S-. R-...Y.F.T  171
Mdo1B        ---RY----- -LTLQT.TLI SVGI.STGV. VTG..LFRQT .K........ MF.L.---.Q T.RT..I.A- -IC..LL--- ------.N-. R-...Y.F.T  171
Sha1B        ---NY----- -LTLQT.TLI SVGI.STGV. VTG..LFRQT .K........ MF.M.---.Q T.RI..I.A- -IC..LL--- ------.N-. R-...YTF.T  171
Aca1B        ---RC----- -LGVRS.SL. S.LI.S.GA. MLGVLLRRQ. .......... MF...---.K TRPS..I.A- -SM..VM--- ------..-. .-...Y.F.T  171
Fpe1B        ---RL----- -LSYES.TMV STVI.STGA. LSAVLLFRQ. .K........ MF...---.K ..RS.RI.V- -..MKVL--- ------.I-. K-.L.Y.F.T  169
Xtr1B        ---SI----- -LAPVS.T.. S.VI.S.GV. VSG.LIYRQT ..I....... MF...---KK T.MK..I.A- -GCMK.M--- ------.S-. Q-...Y.F.M  169
LchB         ----C----- -MSFRS.TL. S.T...TGV. VTG.LAVRYS ..A..C.... MF...---.S TRIR.RI.A- -T..K..--- ------..-. .-...Y.F.T  131
Dm           ---TI----- -----NWQ.T .CF.A.LVV. LSIC..MRYT .K...M.K.. MY.SRAPGSR V.L..ML.V- -.V..VL--- ------.SWN K-.G.Y.F.G  169
ci           ---K.----- -DSKTRN--T SYLVSSTL.. LLAVLVIRYL .K...C.Q.. MF..R---.K ..LK..L.AV -SYLQLL--- ------CYFT Q-NVM.LLVC  169


Hsa1C        SLPRQPVPSV QDTVRKYLES VRPILSDEDF DWTAVLAQEF LRLQASLLQW YLRLKSWWAS NYVSDWWEEF VYLRSRNPLM VNSNYYMMDF LYVTPTPLQA  269
SSc1C        ........A. .......... ...V.CE... E.ISA..R.. .K........ ..Q..Y.... .......... ......S... .......... .N.....V..  269
Mmu1C        A........A .E........ ...V.G.DA. .RATA..ND. ...H.PR..L ..Q....CT. .......... ......GS.- I..T...... ..........  268
Hsa1A        ....L...A. K...NR..Q. ...LMKE... KRMTA...D. AVGLGPR... ..K......T .........Y I...G.G... ......A..L ..IL..HI..  269
Mmu1A        ....L...A. K...SR.... ...LMKEG.. QRMTA...D. AVNLGPK... ..K......T .........Y I...G.G.I. ......A.EM ..I...HI..  269
Mdo1A        ....L...A. K...NR.... .Q.L.NKDN. QRMKG..ED. STNLGPR... ..K......T .........Y I...G.S.I. ......A..L ..IL..TI..  269
Gga1A        ....L...A. K...NR.... ...LMN..E. KRMEG..KD. AFNLGPR... ..K......T .........Y I...G.G.I. .....FA... .HLS..TI..  269
Fpe1A        ....L...A. N...NR.... ...LMD..E. RRMEG..KD. AFNLGPR... ..K......T .........Y I...G.G.I. .....FA... ..LS..T...  269
Aca1A        ....L..... KN..NR.... .H.LMNE.Q. KRMEA.GKD. ATNLGPK... ..K......A .........Y I...G.G.I. .....FA... ..F...SV..  269
Xtr1A        ....L...P. K...KR..D. .K.LMDK.K. ERMEG..KD. ANNLGPR... ..K......T .........Y I...G.G.I. ......A... ...L..HI..  270
Lch1A        ....L...A. K..MTR.... ...LMD..Q. RRMTK..KD. ELKLGRR... ..K......T .........Y I...G.G.I. ......A..L ..........  269
Ler1A        ....L...T. K..M...... ...L.N.KE. QRMQA..KD. ELKVGPR... ..K......T .........Y I...G.G.I. ......A..Y ..IV...V..  269
Xtr1C        V...L...G. KE..QR..D. ...LMN..EY KRMTG..KD. EVNL.PR... ..K......A .........Y I...G.G.I. ......A... .......V..  269
Ler1C        ....L...P. ...L.RF... ...LMN.... RR.KA..KD. E.NL.PR... ..I....... .........Y ...QG.E.I. ......A... ......TS..  269
Cmi1C        ....L...TI ...MQR.... ...LMN..E. HRMEA..KD. EVKLGPR... ..K......T .........Y ....G.E.I. ......A... .......I..  269
Aca1C        ....L...AL K..MQ..... I..LTT.AE. QRM.A..RD. EQTLGPR... ..K....... .........Y ....G.G... ......A... I......V..  269
Lch1C        C...L...PI AK.MQR.... .H.LMD..K. KRMTA..KD. EVNL.PR... ..K....... .........Y I...G.G.I. ......G... I.AS..YI.T  268
Tni1A        ....L...DI S..C.RH... ..ALMD..Q. ERMTA.TKD. EKNLGPR... ..K....... .........Y I...G.G.I. ......A... ...I..SI..  270
Tru1A        ....L...A. S..C.R.... .H.LMDE.R. ERMTA..KD. EKNLGPR... ..K....... .........Y I...G.S.I. ......A... ...F..SI..  270
Oni1A        ....L....I K..CER.... ...LMD.QQ. ERMKG.T.D. EKNLGPR... ..K....... .........Y I...G.G.I. ......A... ...F..SI..  270
Gac1A        ....L....I K..CKR.... ...LME..QY QRMEG.TKD. EKNLGPR... ..K....... .........Y I...G.G.I. ......A... ...F..SI..  270
Dre1Aa       ....L...P. K....R.... A..LMD..QY KRMEG..KD. EKNLGPK... ..K.....T. .........Y I...G.S.I. ......A... ...N..S...  270
Tfu1A2       ....L...P. EH..KR.... ...LMD..QY KRMEA..KD. QSNLGPK... ..K..AF... ........DY ..V.G.G.I. ......A... ...F..H..V  270
Dre1Ab       ...HLF.... KE.T.R.... .Q.L....EH QRMQR..LD. E.NLGPK... ..K......T .........Y I...G.G.I. ......V... ..AF..NI..  271
Tfu1A1       ...NL.L... K..MKR.... ...L.D.TEY KKMEE..SD. QKTL.PK.H. ..K......T .........Y I.......I. ......A... .H.L..H...  271
Ola1Ca       ...NL...AI K...KR.... ...LMN.GEY ERMTK..T.. ESSLGNR... ..K..AL... .........Y ....G.S.I. ......G... .......I..  269
Oni1Cb       ...NL...TI K...KR.... ...LMD.KEY ERMTK..A.. ESSLGNR... ..K..AL... .........Y ....G.G.I. ......G... .......I..  269
Ame1C        ...NL...AI K...KR.... ...LK..SE. ERMTN..K.. EDSLG.R... ..K..AL... .........Y I...G.G.I. ......G... .......I..  269
Gac1C        ...NL...LI K...NRH... ...LMD.TEY ERMTS.SED. ESGLGKR... ..K..AL..T .........Y ....G.S.I. ......G... M......I..  271
Tru1C        ...NL...A. K...KR.... ...LMD.AQY EHVTK..A.. ESSLGNR... ..K..AL.VT .........Y ....G...I. ......V... .......I..  269
Tni1C        ...NL...A. K...KR.... ...LMD.AQY ERV.K..A.. ESSLGNR... ..K..AL.VT .........Y ....G.S.I. ......V... .......I..  269
Ola1Cb       ...HL...A. R..LTR.... ...L.T.PEY KRMTD..N.. ESSLGNR..R ..K..AL..T .........Y I...G.G.I. ......G... ......SV..  271
Oni1Ca       ...HL...AI ....SR.... ...L.T.LE. KRMTD..N.. ESNLGNR..R ..K..AL..T .........Y ......G.I. ......G... ......SV..  269
Gac1C        ...HL...PI K...SR..T. A..L.T.PE. ERMTK..GQ. EANLGNR..R .....AL..T .........Y I...G.G.I. ......G... .......V..  270
Tri1C        ...HL...AI K..LSR..R. ...L.N.LEY KRMSE..SD. EKNLGNR..R ..K..AL..T .........Y I...G.G.I. ......G... ......SV..  271
Dre1C        ...HL...PI K..LER.... .K.L.DLDG. QRMRR.TS.. EKSLGNR..R .....AL..T .........Y ......S.I. ......G... .......N..  269
Dre1B        ...HL....I D..I.R.... ...L.D..QY KQMETV.ND. KKDP.PK..K H.K......T .........Y I...G.D.I. ....F.T..L ...I..YR..  269
Oni1B        ....L..... D..IHR.... ...L.DN.QY NKMEL..SD. KENK.AQ..R C.I......T .........Y I...G.S.I. ....F.I..L ......HR..  269
Gac1B        ....L..... D..IHR.... .H.L.NSDQY .QMER..ND. KDSK.AQ..R ..I.....GT .........Y I...G.S.I. ....F.I..L ..I...HR..  269
Ola1B        ....L...R. D..I.R.... ...L.DK.QY SQMET..ND. KESK..Q..R ..I......T .......... I...G.G.I. ....F.I..L ..I...HR..  269
Tru1B        ....L..... D..IHR.... ...L.VSDEY .QMVT..K.. KDSK.AQ..R ..I......T .........Y I.....S.I. ....F.I..L ..I...HR..  269
Tni1B        ....L..... D..IHR.... ...L.VSGEY NQMVA..N.. KDSE.AQ..R ..I......T .........Y I.....S.I. ....F.I..L ......HR..  271
Hsa1B        ...KL...R. SA.IQR.... ...L.D..EY YRMEL..K.. QDKT.PR..K ..V....... .........Y I...G.S... ......V..L VLIKN.DV..  271
Mmu1B        ...KL..... PA.IHR..D. ...L.D..AY YRMET..K.. QDKT.PR..K ..V......T .........Y ......S... ......A... VLIKN.NV..  271
Mdo1B        ...KL...K. AA..HR.... ...L.D..QY YRMEM..KD. QEKT.PR..K ..V......T .........Y I...G.S..V ......V... VFTQH.NI..  271
Sha1B        ...KL...K. SA.IKR.... ...L.D..EY YRMEM..KD. QERT.PR..K ..V......T .........Y I.....G.I. ......V... VLTPH.EV..  271
Aca1B        ...KL...R. A..IQR.... ...L.DE.R. LDMEA..LD. QQRL-IR..K ..L......T .........Y I...G.S.I. ......V... ..T...H...  270
Fpe1B        ...KL...P. EA.ITR.... ...LMD..KY SKMEA..K.. KEKT.PR..K ..I.....TT ........QY I..HG.S... ......A... .....SHI..  269
Xtr1B        ...KL...PL E..IER..Q. ...L.D.DK. SEMKI..E.. QKDLGRK..K ..H......L .........Y I...G.G.I. ......A..Y .....STN..  269
LchB         C..HL..... E..LHR.... ...L.N.LQY KRMEA.TIQ. KHQT.PR..K ..L......T .........Y ......T.I. ......A..L ..II.SSV..  231
Dm           ....L.L... K..MTR..R. ...L.D..NY TRMER..K.. EQTIGKK... ..I.....ST .........Y ....G.S..C ....F.GT.A IFMNL.DK..  269
ci           LITKTNIIK. EGLGI..YKE LKRFIYEKSY EKNVLRSLH. TSN------K L.ITNPDFI. IHS.NR..QY ...AG.G.I. ......G..L ..HN...V..  263


Hsa1C        ARAGNAVHAL LLYRHRLNRQ EIPPT-L--L MG-MRPLCSA QYEKIFNTTR IPGVQK--DY IRHLHDSQHV AVFHRGRFFR MGTHSRNSLL SPRALEQQFQ  363
SSc1C        .......... .......... ..F..-.--. ..-....... .......... ....HR--.H .H..R..R.. .......... V....QSG.. ..........  363
Mmu1C        ........T. .....L.... ..S..-.--. ..-....... ...RM..... ....E.--.H L...Q..R.. .......... V....P.G.. ..........  362
Hsa1A        ......I..I ....RK.D.E ..K.I-R--. L.STI..... .W.RM...S. ...EET--.T .Q.MR..K.I V.Y....Y.K VWLYHDGR.. K..EM...M.  364
Mmu1A        .....TI..I ....RTVD.E .LK.I-R--. L.STI..... .W.RL...S. ...EET--.T .Q.VK..R.I V.Y....Y.K VWLYHDGR.. R..E....M.  364
Mdo1A        ......I.SI ....KK.D.E .LK.I-.--. L.STV..... .W.RM...S. ...EET--.T .Q.VK..R.I V.Y....Y.K VWLYHDGR.. K..EI...M.  364
Gga1A        .....II..I ....KK.D.. ..K.I-.--. ..STV..... .W.RM...S. ...EES--.T LQ.VK..K.I V.Y.K..Y.K VWLYHDGR.. K..EI...I.  364
Fpe1A        .....VI..I ....KK.D.. ..K.I-.--. ..STV..... .W.RM...S. ....ES--.V LQ.VK..K.I V.Y.K..Y.K VWLYHDGR.. K..EI...I.  364
```

```
Aca1A      ......I..I ....RK.D.. Q.Q.I-.--. ..STV..... .W.RM...S. ...EET--.T .Q.IK..K.I V.Y.K.C.YK VWLYYDGR.. K..EI...M.  364
Xtr1A      ......I..I ....RK.D.E ..K.I-Q--. ..STV..... .W.RMY..S. ...EET--.T .Q.VK..K.I V.Y.K..Y.K VWLYHDGR.. K..EI.H.M.  365
Lch1A      .....TI..I ....RK.D.E ..K.L-M--I QN-TI.M..S ...RM...S. ....ET--.T LQ.MR..K.I V.Y.K..Y.K VWLYHDGR.. K..EI...M.  363
Ler1A      .....TI..M ....RK.D.E ..H.L-M--- -.-QI.M..S ...RM..SS. ...LET--.T MQ..K..K.. V.Y.K..Y.K VWLYHSGR.. ..CEIQL.M.  361
Xtr1C      ...A.T...I ....RKVT.E .LK.L-M--I QD-CL.M... ...HM..... ....DT--.T .Q.YA..K.I V.Y.K..... VWVYQGGR.. N.KE..L...  363
Ler1C      .....L.... ....R..A.E QVQ.S-T--. P.FPI..... .W.RM..S.. L..EET--.R LQ.QA..K.. M.Y.K..Y.K VWLYQSGR.Q ..SE.QK...  364
Cmi1C      .....LT... ....RK.T.E ..K.S-T--. P.LPV....S .W.RM..S.. T..QET--.H LQ.QT..K.I V.Y.K..Y.K VWVYQGGR.. R..E..V...  364
Aca1C      .....L.Y.M MM..RK.VQE ..K.M-.--I Q.-CL.M... .W.RM..... ...MEG--.S .Q..Q..R.I ..Y.A..... VLLYHNGRM. R..E.QA...  363
Lch1C      ......IY.C .Q..RKVT.E .LK.L-.--I QD-CL.M... .H.HM...S. ...IET--.T LQ..T..K.I V.Y.K..... VWVYHGGRM. K..E..I.IE  362
Tni1A      ......I..I M...RK.D.A Q.K.L-M--. QN-TI.M... ...RM..... V...ET--.T LQ.TNETK.I V.Y.K....K VWMFYDGR.. L..EI...ME  364
Tru1A      ......I..I M...RK.D.A Q.K.IY.--. AN-KV..... .W.RM..... V...ET--.T LQ.TIETK.I V.Y.K....K VWVFYDGR.. L..EI...ME  365
Oni1A      ......I.SI M...RK.D.A Q.K.L-M--. LH-TI.M... ...RM..... V...ET--.T LQ.VNE.K.I V.Y.K..Y.K VWMFYDGR.. L..EI...ME  364
Gac1A      ......I..I M...RK.D.A Q.K.IY.--. AN-KV..... .W.PM..... V..LET--.I LQ.VN..K.I ..Y.K....K VWMFYDGR.. L..EI...MA  365
Dre1Aa     .....SI..M MM..RK.D.A Q.K.L-M--V LN-TI.M..S ...RM...S. V...ET--.V LQ.VNE.K.I ..Y.K...YK VWMFYDGR.. L..EI...ME  364
Tfu1A2     ......I.SI M...RK.D.A Q.K.L-M--V QN-TI.M..S ...RM..... V...EE--.F .K.VNE.K.I V.Y......K VWMFYDGR.. L..EI...ME  364
Dre1Ab     .....VI..I M...RK.D.A Q.K.L-M--. QN-TI.M..S ...RM...S. ...IET--.S VQ.VS..R.I V.Y....Y.K VWMFYDGR.. L..EI...ME  365
Tfu1A1     .....TI.SI M...RK.D.A Q.K.L-.--. QN-TI.M..S ...RM...S. ...IET--.T .Q.VS..R.I V.Y.K..Y.K VSMFYDGR.. L..EI...IE  365
Ola1Ca     .....SI..F F...RK..KE ..K.S-R--I P.TVI...A. .C.R...... ...EET--.T VQ.WK..DY. ..Y.K..Y.. LRVYQAGR.. ...EI.F.I.  364
Oni1Cb     .....SI.SF F...RK..KE .LK.W-.--. RS-AV.C..Y .F.RM.D.C. ...ILTAK.T VQ.WQ..DYI V.Y.K..Y.. LRVYQAGR.. ...EI.F.I.  365
Ame1C      .....TL..V .M..R...KE ..K.S-R--I P.TFI...A. .C.R...... ...EET--.T VV.WQ..EY. ..Y....Y.. LWLYQAGRM. ...EI.Y.I.  364
Gac1C      .....SI..Y F...RK..KE .LK..-R--I P.TVI..... .C.RM..... ...EETGK.T VQ.WQ..DYI ..Y....Y.. LRMYHAGR.. ...EI.S.I.  368
Tru1C      .....TIY.M ....CK..KE ..K.P-AQW. LRSAV.C..Y .F.RM...C. ...TLT--.T .Q.WK..DFI V.Y.K..Y.Q LYVYQDGR.. C..EI.F.I.  366
Tni1C      .....TIY.M ....SK..KE ..K.P-AQW. LRSAV.C..Y .F.RM...C. ...TLT--.T VH.WN..ECI V.Y.K....Q LCVYQEGR.. C..EI.F.I.  366
Ola1Cb     .....TIT.. ....RMV..E .LT.S-R--V P.TVI...A. .C.RM..... T...ET--.V LQ.WQ..EF. ..YS...YY. LWVYRAGR.. .A.EI.H.I.  366
Oni1Ca     .....TIT.. ....RKV..E .LK.S-R--V P.TVI...A. .C.RM..... T...ET--.V LQ.WL..EF. ..Y....Y.. LWVYRAGR.. ...EI.Y.I.  364
Gac1C      .....TIT.. ....RKV..E .LK.S-R--V P.TVI...A. .C.RM..... T...ET--.V LQ.WQ..EF. ..Y....... LWVYLAGR.. ...E..H.I.  365
Tri1C      .....TIT.. ....RKV..E .LK.L-W--C ICTVI...A. .C.RM..... T..EET--.V LQ.WQ..EF. ..Y.K..... LWVYRAGR.M ...E..Y.I.  366
Dre1C      .....TIT.. F...RKV..E .LN.S-R--I P.TVI...A. .C.RM..... T..EET--.V LQ.WQ..EF. ..Y....Y.. LWVYRAGR.. ...EIQF.I.  364
Dre1B      .....V...M .Q..RK.E.G .LT.L-R--A L.-IV.M..F ...RM..... ...IET--.F VQ..K.RK.L V.Y....L.K VWLYYGGRH. W.SE..L...  363
Oni1B      .....V...M .Q..RK.E.G .LA.L-R--A L.-TV.M..T .M.RM..... ...IET--.F VQ..T.RK.L V...K....Q VWLYTGGRH. L.SE..T...  363
Gac1B      .....V...M .Q..RK.E.G .HA.L-R--A L.-TV.M..T .M.R...... ...IET--.I VQ..T.RK.L V.Y.K....L LWLYTGGRH. L.SE..T...  363
Ola1B      .....I...M .Q..RK.E.G .HA.L-R--A L.-TV.M..T .M.RM..... ...IET--.V VQ..S.RK.L I.Y.K....Q VWLYTGGRH. L.SE..M...  363
Tru1B      .....M...M .Q..RK.E.G .HA.L-R--A L.-TV.M..T .M.RM..... L..IET--.A VL..T.RK.L I.Y.K....Q VWLYTGGRH. L.SE..L...  363
Tni1B      .....T...M .Q..RK.E.G .HA.L-R--A L.-TV.M..T .M.RM..... ...IET--.V VQ..T.RK.L I.Y.K....Q VWLYTGGRH. L.SE..L...  365
Hsa1B      ..L..II..M IM..RK.D.E ..K.V-M--A L.-IV.M..Y .M.RM..... ...KDT--.V LQ..S..R.. ..Y.K....K LWLYEGAR.. K.QD..M...  365
Mmu1B      ..L......M IM..RK.D.E ..K.V-M--A L.-.V.M..Y .M.RM..... ...KET--.L LQ..SE.R.. ..Y.K....K VWLYEGSR.. K..D..M...  365
Mdo1B      ..L.SV...M IM..RK.D.E ..K.V-M--A L.-IV.M..Y .M.RM..... L..KDT--.V LQ..L..R.. ..Y.K...YK LWLYQGSQ.. K..D..M...  365
Sha1B      ..L..V...M IM..RK.D.E ..K.V-M--A L.-IV.M..Y .M.RM..... ...KES--.V LQ..V..R.. ..Y.K...YK VWLYQGTQ.. K..D..M...  365
Aca1B      ........SI ....R..D.E DLA.V-M--A L.-VV....Y .M.R...... ...K.A--.R LL..S..R.L ....K....K VWLYHAGK.. P..D..M...  364
Fpe1B      .....M...I .M..RK.D.G ....M-M--A L.-IV.M..Y .S.RM..... ...KET--.T LL..V..K.L ..Y.K...YK VWLYYGGQ.. Q.CD..L...  363
Xtr1B      .....VI..M ....RK.E.G L...V-M--A L.-IV.M..N .MVRM..... V...ET--.C LQ..VE.R.. C.Y.K..YY. LALYENGN.. T..Q.QA.I.  363
LchB       .....T...M ....RK.D.E ..K.M-M--A LK-LV.M..N .V.RM..... ...LET--GT VK---Q.AL. YIYNKAV.VD CMLY.LS--. VKLDIL.KQK  320
Dm         ...A.VISL. .NF.RLIEH. .LQ.I-M--V Q.-.I....W ...RT...A. V..LET--.R .I.YR..N.I V.L.K.CYYK .LIYYKGRI. R.CE.QV.IE  363
ci         S..A.I...M FAF.ST.DKE L.K.I-K--V N.-VV..... ...RV..SC. T...EA--.R .C.WS.CR.I ..Y.Q..W.K .TCYKNGV.. E.SEM.I.IE  357


Hsa1C      RILDDPSPAC PHEEHLAALT AAP----R-- GTWAQVRTSL KTQA--AEAL EAVEGAAFFV SLDAEPAGLT -------RED PAASLDAYAH ALLAGRGHDR  448
SSc1C      .......... .......... ...----.-- DM.....K.. ....--E... .......... ...S....DA AGDTPEPSG. S......... ..........  455
Mmu1C      D......... .L........ ...----.-- SM.....E.V ..H.--.T.. .......... ...S...... -------... .......... ..........  447
Hsa1A      ....NT.EPQ .G.AR..... .GD----.-- VP..RC.QAY FGRGKNKQS. D...K..... T..ETEE.YR -------S.. .DT.M.S..K S..H..CY..  451
Mmu1A      Q....T.EPQ .G.AK..... ..D----.-- VP..KC.QTY FARGKNKQS. D...K..... T..ESEQ.YR -------E.. .E..I.S..K S..H..CF..  451
Mdo1A      .......EPQ AG..K..... .GD----.-- VP..KA.QTY FSRGKNKQS. D...K..... TM.DTEQ.YS -------KK. .LT.M.S..K S..H.KCY..  451
Gga1A      .....D.EPQ AG..K..... .GD----.-- VP..KA.QAY FSRGKNKQS. D...K..... T..DDEQ.YS -------K.. .VS......K S.IH..CY..  451
Fpe1A      ...N.D.EPQ AG..K..... .GD----.-- VP..KA.QAY FSRGKNKQS. D...K..... T..DIEQ.YR -------KD. .VK......K S.IH..CY..  451
Aca1A      W....K.KPQ .G..K..... .GD----.-- VP..KA.QTY FARGKNKQS. D.I.K..... T..DTAQ.YR -------E.. .VTTMET..K S..H.KCY..  451
Xtr1A      K.I..T.SPQ .G..K..... .GD----.-- VP..KA.KAY FANGKNKQSM D...K..... T..ETEQ.YN -------K.. .VN...S..K S..H.KCY..  452
Lch1A      .....D.KPQ .G..K..... .GD----.-- VP..KA..TY FCRGKNKLS. D...K..... T..DTEQ.FR -------K.. .VT...R..K S..H.KCY..  450
Ler1A      K....Q.LPQ TG..K..... .GD----.-- VP..KA.Q.Y FS.GRNKLS. ..I.K..... T..DTEQ.FR -------K.E .IS...N..K S..H.KCY..  448
Xtr1C      N..N....PQ .G..K..... .GE----.-- TA..KA.KTY FRSGKNLQ.. DL..R..... T.QDDEE..R -------T.. .VN.....GK S..H.KCY..  450
Ler1C      Y....A.IPQ .G.KY..... .GK----.-- IP..K..K.Y FSSGKNKT.M DS..K..... T..EDTPE.F -------VDN QVK...Q..K S..H.KCY..  451
Cmi1C      ...A.T.LPQ .G........ .GN----.-- IP.GKA.K.F FSNGKNRSS. DC..K....L T..GDKP..Q -------V.. .VK......K L..H.KCY..  451
Aca1C      S..G..T.PS .G..K.P... .GE----.-- DP..RA.NAF FQTGQNEQS. SI..K..... T..TSEQ..R -------EPN .GQ......K S..H..CC..  450
Lch1C      K..A.K.SPL .G....P... .GD----.-- VP..KA.RDY FQSGLNRQS. DL..K..... ...ESEQ..K -------TD. ..K......K L..H.KCY..  449
Tni1A      K..A.Q.APQ .G..K..... .GTG---.-- TP..NA.DTY FSRGKNKQ.. D.I.K..... T..DTEQRYD -------TNN .VV...S..K C..H.KCY..  452
Tru1A      ...A.Q.EPL .G..R..... .GD----.RG TP..NA.DTY FSRGKNKQS. D.I.K..... T..DTEQCYD -------TNN .VT...S..K S..H.KCY..  454
Oni1A      ...A.K.EPL .G..R..... .GD----.-- TP..KA.E.F FSRGKNKQS. D...K....L T..DTEQRYD -------TKN .VK...I..K S..H.KCY..  451
Gac1A      ...A.TTEPM .G..K..... .GD----.-- TP..NA.ETY FSRGKNKQS. D.I.K...C. T..DTEQRFE -------SDN .DQ..VS..K S..H.KCY..  452
Dre1Aa     ...A.K.EPQ .G..F..... .GD----.-- VP..KA.SQF FIRGKNKQS. D...K..... T..DSEQRYE -------PDN .IQ...S.GK S..H.KCY..  451
Tfu1A2     ...A.T.MPQ .G..T..... .GD----.-- VP..KA..EF FSTGKNRKS. D...R..... T..DTEQRYE -------PDN .VQ...S..K S..H.KCY..  451
Dre1Ab     ...A.T.EPQ .G..T..... .GDSVCQ.-- VP..CA.NAY LRHGTNKKS. DS..K..... T..DTEQRFD -------QKN .VE...R..K S..H.KCY..  456
Tfu1A1     ...A.T.EPQ .G..K..... .GD----.-- VP..CA.DAY LR.GKNRQS. D...K..... T..DTEQRHN -------SDS ..E..RSFGK S..H.KCY..  452
Ola1Ca     ......A.PS KG.AK.G... ..D----.-- VS..EA.VKY FSSGINKRS. DVI.R..... T..D.EQ.TM -------.D. QP....N..K S..H.KCY..  451
Oni1Cb     ........PS KG.AK.G... .GD----.-- IP..KA..KY FSSGVNKRS. DCI.K..... T..D.EQ.MM -------GD. ......R..K S..H.KCY..  452
Ame1C      ........PA .G..K.G.F. .GD----.-- IP..KA.KEF FSSGVNKRS. DCI.K..... T..DDEQ.MM -------GD. ...NV.R..K S..H.KCY..  451
Gac1C      K.......PS KG.AK.G... .GD----.-- IP..KA.AKY FSSGVNKRS. DFI.K..... T..DDEQ.GV -------AD. .-TRI.S..K S..H.KCY..  454
Tru1C      .........S KG.AR.G... .GD----.-- IP..KA.AKH FNSGINKKS. DCI.K..... T..D.EQSIV -------GDN LGE...C.IK S..H.KCYN.  453
Tni1C      .......APS KG.AK.G... .GD----.-- TP..RA.AKY FSSGVNKKS. DCI.K..... T..D.EQ.IM -------GDN LRE...H.IK S..H.KCYK.  453
Ola1Cb     W.......PL .G..K.G... .GD----.-- VP...I.KEH FSSGVNKRS. DII.K..... T..D.AQ.MK -------GD. .TGN..R..K S..H.KCY..  453
Oni1Ca     .........L .G..K.G... .GD----.-- IP...M.KQY FSSGVNKRS. D.I.R..... T..D.EQ.MR -------GD. .EGN..S..K S..H.KCY..  451
Gac1C      ........PQ .G..K.G... .GESFRP.-- VP.F.M.ERH FSSGINKRS. DCI.R..... T..D.EQ.MR -------G.. ..GN..R..K S..H.KCY..  456
Tri1C      K.......PQ .G..K.G... .GDR---.-- IP.....KQY FSSGVNKRS. DVI.K....I T..D.EQ.MR -------G.. ..GN..R..K S..H.KCY..  454
Dre1C      ........PS .G..K.G... .GN----.-- TP..R..KQF FSSGVNKQS. DCI.K..... T..DQAE.MK -------G.N .SEN..R..K S..H.KCY..  451
Dre1B      .....K.EPQ .G.LK.PS.. .GN----.-- VP..RA.LKY FGEGVNRAS. ..I.T....L T..D.AH.YD -------P.N I-R...L..K S..H.KCY..  449
Oni1B      ...N.T.EPQ .G.LK..... .GY----.-- IP...A.IKY FS.GINKVS. D.I.S....L T..D..Q.YD -------PAK S-N...S..K S..H.KCY..  449
Gac1B      ...N.T.EPQ .G.LK..... .GH----.-- VP...S.IKY FG.GVNKVS. D.I.S....L T..D..Q.YD -------PAK A-K...S..K S..H.KCY..  449
Ola1B      ...N.TTEPQ .G.LK..... .GN----.-- VP...A.IKH FSHGVNKTS. D.I.S....L T..D.SQ.YD -------GVK K-N...S..K S..H.KCY..  449
Tru1B      ...N.T.EPQ QG.LK..... .GN----.Q- VP..RA.SKY FS.GLNKVS. D.I.S....L T..D..Q.YD -------QAR .-R...S..K S..H.KCY..  450
Tni1B      ...S.T.EPQ QG.LK..... .GN----.-- VP...A.AKY F..GLNKAS. D.I.S....L T..D..Q.YD -------HAR S-R...S..K S..H.KCY..  451
Hsa1B      ........PQ .G..K..... .GG----.-- VE...A.QAF FSSGKNKA.. ..I.R..... A..E.SYSYD -------P.. E-...SL.GK ...H.NCYN.  451
```

```
Mmu1B         ........PQ .G..K..... .GG----.-- VE..EA.QTF FSSGKNKMS. D.I.R..... T..EDSHCYN -------PD. E-T..SL.GK ...H.NCYN.  451
Mdo1B         ........PQ .G..K...F. .GG----.-- VQ..EA.QTY FNTGKNKAS. ..I.K..... T..E.SH..D -------P.N E-...SL.GK S..H.NCYN.  451
Sha1B         .....TC.IQ .G..K..... .GG----.-- VQ..EA.QTY FNTGKNKAS. ..I.K..... T..E.SH.YD -------P.. E-...SL.GK ...H.NCYN.  451
Aca1B         ........PE TG..R..... .GE----.-- LP..EA.EKY FSRGKNKAS. DC..R..... T..E.EH.FD -------PDK E-D...R.SK S..H.QCC..  450
Fpe1B         ........PQ .G..R..... .GE----.-- VP..EA.ARF FSHGKNKVS. D.I.R....L T..E.EH.YV -------AGK E-GCM.T..K S..H.QCY..  449
Xtr1B         Y....S..PQ .G..K..... .GN----.-- VH...A..NF FSNGINRT.. SC..R.V..I ...E.E..YN -------E.. K-S..S..SK ...H.NCYN.  449
LchB          ..KIIT.YIK YERKGSVRIS LSC----.-- IP..KA.SEF FSHGKNKIS. T.I.R....M T..D.EQAYD -------K.N .-TT..S..K S..H.KCF..  406
Dm            E..KGKATPV EG........ .WN----.-- SK..EA.NTF FSWGVNQTS. RTI.S...VL ...D..FEFD -------LAR .-EL..NFGK K..H.N.YN.  449
ci            S..N.T..P. EG........ .GE----.-- IP..KA.NTY FVDGVNKKS. H.I.K...IL V..D.EHVVS -------D.. S-...SK.GR S..H.KCYN.  443


Hsa1C         WFDKSFTLIV FSNGKLGLSV EHSWADCPIS GHMWEFTLAT ECFQLGYSTD GHCKGHPDPT LPQPQRLQWD LPDQIHSSIS ----LALRGA KILSENVDCH  544
SSc1C         ......S..I .......... ..A....... .......... ........A. .........S .......H.. ...K..L... ----...... QA.A..I...  551
Mmu1C         .......... .......... ........V. ..L....... .......A.. .......... .......... ..E..QP... ----...... .T..G.I...  543
Hsa1A         .......FV. .K...M..NA ......A..V A.L..YVMSI DSL....AE. .....DIN.N I.Y.T..... I.GECQEV.E ----TS.NT. NL.AND..F.  547
Mmu1A         .....I.FV. .K.S.I.INA ......A..V ..L..YVM.. DV......E. .....DKN.N I.K.T..... I.GECQEV.E ----TS.SS. SF.AND..L.  547
Mdo1A         ....T..F.. .K...M..NA ......A..V ..L..YVM.. D......TE. .....DTN.N I.Y.T....E I.EECQDV.E ----ES.SL. ST.AND..F.  547
Gga1A         ....T...V. .K...M..NA ......A..V ..L..NVM.. .YLE...LE. .....DTNQN I.I.TK...E I.EECQDV.E ----RS.ST. RA.ADD..FY  547
Fpe1A         ....T..... .K..RI..NA ......A..V ..L..NVM.. .YLE....E. .....DINQN I.I.TK...E I.AECQEV.E ----RS.ST. IA.ADD..FY  547
Aca1A         .......... .K...M..NT ......A..V ..L..NVMFS D.LE...TE. .....ESSSG ILM.S....E ILEECQEV.E ----RS.AV. RP.ADD..F.  547
Xtr1A         ....TMSFV. .K...M.MN. ......A..V ..L..YVM.. DKME...NE. .....DVNGN I.P.S..... I.EECQNVVE ----ES.TV. .A.ADD..F.  548
Lch1A         .....LSFV. .K...M..NS ......A..V ..L..YV... DS.....TEE .....E.K.S I.F....R.E I.EECQEV.E ----IS.KV. .A.ADD..F.  546
Ler1A         .....LSF.I .K...I..NA ......A..I ..L..YV... DQ.....TDE .N...E.N.Q IQP....... IEEPCQEV.H ----QS.SV. QQ.ADD..F.  544
Xtr1C         .......F.. .K...I..NA ......A..V ......V... D......NEE .N...QV.SN ..V......E ISEECQEV.Q ----SS.AV. QA.ADD..F.  546
Ler1C         .......F.. .A...V..NA ......A..I ..L..YA... DT.....KE. .N...D.A.N V.L....... I.KECQEV.M ----SS.KV. QT.AND..FY  547
Cmi1C         .......... .E...V..NA ......A..I ..L..YV... DS.....NDQ .....DAES. VLS....... I.EACQEV.. ----GS.KV. QS.AND..F.  547
Aca1C         .......... YR...S..NA ......A..V ..L..YC... DA.T...DAY .N...DM..N V.P..K...E I.PECEAV.M ----QSF.V. YN.ASDI.F.  546
Lch1C         ........V. .K......NA ......A..I ..L..YV... DT..I..KP. .....EA.S. ILP....... I.AECREV.K ----MS.AV. QT.AND..F.  545
Tni1A         ......N... .K..TM..NA ......A..V ..L..QV.SM DPKN...TE. ...R.A.H.N ..G....... ISTECQQV.Q ----SS.TV. QK.ADD..S.  548
Tru1A         ......N... .K..TM..NA ......A..V ..L..HV.SM DPNN...TEE ...R.V.H.N ..G..K.... I.AECQQV.Q ----NS.TV. QN.ADD..S.  550
Oni1A         .....LNM.. YK..TM..NA ......A..V ..L..HV.SM DP-K...TE. ...V.K.H.N ..G......T I.AECQEA.E ----SS.TV. RA.ADD....  546
Gac1A         ......N..I YK..TM..NA ......A..V ..L..HV.SM DPIK...TEA .....E.H.N ..G..K.S.. I.AECQEV.Q ----SS.KV. RT.ADD..S.  548
Dre1Aa        .....LN... .K..TM..NA ......A..V ..L..QV.SS DPVR...TEE .....N.H.N M.G....... I.EECQTV.. ----SS.KV. NT.ADD..M.  547
Tfu1A2        ......N..I .K..TM..NA ..T...A..V ..L..HV.SM DPIT...TE. ...R.K.H.N ..G.L..... ISVECQ.V.R ----SS.KV. NA.ADD..M.  547
Dre1Ab        .....IN..I .K..TM..NA ......A..V ..L..QV.SM DPVK...TE. .....E.HAN ..G......N I.TECQTM.T ----NS.SV. EA.ADD..S.  552
Tfu1A1        ......N..I YK.ATI..NA ......A..I ..L..NV.S. DAIK...TD. ...A.QTHRN ..G......N I.PECQTM.A ----NS.SV. QA.ADD..MV  548
Ola1Ca        ......SV.Y YK...S.ING ......A.VV A.V..YV... DS.....NEE .....EV.AS .....K.N.E ISPECEEQ.. ----RS.AV. QA.ADD..F.  547
Oni1Cb        .......VVY YK...N.INA ......A.VL A.V..Y...N DS.....NAE .....DV..S ..R.VK.S.E I.PECEEQ.A ----QS.AV. QA.ADD..F.  548
Ame1C         ......SVVI .K...N..NA ......A..V S....YA... DS.....NEE .....DVN.S .......T.. I.K.CQEQVA ----QC.AV. QP.ADDI.F.  547
Gac1C         ......SVVY .K...M..NG ......A.VL S.A.QYV.T. D......NAE .....EV.SS ..G..K.N.E I.PECEEQ.. ----GS.AV. QA.ADD..V.  550
Tru1C         ......SVVF YK...S..NG ....G.A.VL T.L..Y.... .......NAE .....EV.AS .AE..K.N.E ISSECEEQ.. ----QS.EV. QA.AND..M.  549
Tni1C         ......SVVF YK...S..NG ....G.A.VL S.L..Y.... .......NAE .....EV.AS ..K..K.N.E I.SECEEQ.C ----QS.AL. QA.AND..M.  549
Ola1Cb        ......SIVI YK...S..NA ......A.TV A.L..Y.... DA.....TE. .....EV.RS ..P.H..S.E I.SEVQDQ.F ----SS.TL. .A.ADD....  549
Oni1Ca        ......SVVI YK...S..NA ......A.TV A.L..Y.... DA.....TE. .....EVE.S .......V.N I.AEVQAQV. ----SS.AV. QA.ADD....  547
Gac1C         ......SIVI YK...S..NA ......A.TV A.L..Y.... DA.....TE. .....DV.RS ..P....A.. I.SEVQAQA. ----SS.VV. QA.ADD....  552
Tri1C         ......SIVI YK...N..NA ......A.TV A.L..Y.... DA.H...TE. .....EVE.L ..H....L.. I.LECNTQVQ AQQCSS.AV. QA.ADD....  554
Dre1C         ......SVV. YK...N..NA ......A..V A.L....... DT.H...NS. .N.R.DV.HS ..H....S.. I.FEVQTQ.. ----ES.AV. QA.ADE....  547
Dre1B         ......N... YK...M.VNT ......S..I .....YV... D..H...TAE .....DVNK. .AP.T..... I.KACQEI.E ----GSY.I. .GIADD..F.  545
Oni1B         .........S YP...M.VN. ......A..V .....YI... D..H...TEE .....DVNKN ..H.T....Q I.NECQNV.E ----TSYLS. .QIADD..F.  545
Gac1B         .........S YP...V.VNA ......A..V .....YV... D..H...TEE .....DVNKG ..H.S....Q I.NECQEV.E ----TSYLS. .LIADD..F.  545
Ola1B         .........S YP...M.VN. ......A..V .....YV... D..H.C..EE .....DANRG ..F.T....Q ISKECQDV.E ----ASYLS. .KIADD..F.  545
Tru1B         .........S YP...M.IN. ......A..V .....YV.S. D..H...TEE .....DVNKG ..Y.S....Q I.VECK.I.E ----ASYVS. .RIADD..FY  546
Tni1B         .........S YP...M.IN. ......A.VV ......V... D..H...TEE .....DVNKG ..Y.S....Q I.VECQ.I.E ----ASYVS. .QIADD..F.  547
Hsa1B         .........S .K..Q...NA ..A...A..I ..L...V.G. DS.H...TET ...L.K.N.A .AP.T..... I.K.CQAV.E ----SSYQV. .A.ADD.ELY  547
Mmu1B         .........S CK..L....NT ......A..I ..L...V.G. DT.H...TET ...V.E.NT. ..P....P.. I.E.CREA.E ----NSYQV. .A.ADD.ELY  547
Mdo1B         ......N..S .K.A....NT ......A..V ..L...V... DA.H.D.TDA ...Q.K.NHS .AP....L.. I.EECQEL.E ----SSYQV. .T.ADD.ELY  547
Sha1B         ........VA .K......NT ..A...A.VV ..L...V... DA.H.D.NES ...Q.K.NHS .AP......E I.EECQKI.E ----SSYEV. .A.ADD.ELY  547
Aca1B         ......S.V. YR.....ANA ......A..I ..L...M... DH.....CS. ...H.V.NTA ..P....T.. I.EECCNV.D ----ASYAV. RA.ADDI.F.  546
Fpe1B         ........V. YK.....ANA ......A..I ..L...A... .K.....TDR ...R.E.NTQ .AP....... I.QECRDT.E ----SSY.L. .A.ADD..FC  545
Xtr1B         ......S.V. .R......NA ......A..I ..L....... D..E...TE. .N.R.DAGSP ..P.Y..... I.PKCREV.E ----RSYVT. .AIADD..F.  545
LchB          .......F.. .K.....INT ......A.VI ..L...V... D..E.S.TES .....EMNKK ..P....... ..EECKEM.Q ----QSYKV. .A.ADD.NFC  502
Dm            ....C..VC. GT..RV.FNA ..T.S.AA.A S....NLIVD DLVSD..DET .NT..T.AFQ P.T.T..T.. .KP-CLAQ.E ----E.TIDV TK.INE.NLR  544
ci            ....T.NC.. .K..RW.INA ......A..M SYVV.EA.GF .YQS...TQ. .RV..R.TVQ PIT.H....Q .TPECQEV.E ----TS.SV. NN.ADD.HLN  539


Hsa1C         VVPFSLFGKS FIRRCHLSSD SFIQIALQLA HFRDRGQFCL TYESAMTRLF LEGRTETVRS CTREACNFVR AMEDK---EK TDPQCLALFR VAVDKHQALL  641
SSc1C         .F...H.... ..K....... ....T..... ......R... .......... .......... ......S... ...HQ---.. .......... L.........  648
Mmu1C         .F...H.... ..KC..V... ....LV.... .......... .......... .......... ......Q... ..DN.----E ..QH...... ..........  639
Hsa1A         SF..VA...G I.KK.RT.P. A.V.L..... .YK.M.K... ...AS..... R......... ..T.S.D... ..V.P---AQ .VE.R.K..K L.SE...HMY  644
Mmu1A         SF..DT...G L.KK.RT.P. A...L..... .YK.M.K... ...AS..... R......... ..T.S....L ..M.P---TT .AE.RFK..K I.CE...H.Y  644
Mdo1A         SF..DA...E L.KKSRT.P. A.V.L..... .YK.M.K... ...AS..... R......... ..M.S....L ..VNP---TE SVENK.K.L. I.AE...HMY  644
Gga1A         SFY.DV...G L.KKAKT.P. A...L..... .Y..M.K.S. ...AS..... R......... ..I.S....Q T..NP---SE SNENKMKS.. L.AT...H.Y  644
Fpe1A         SFF.DA...G L.KKAKT.P. A.V.L..... .Y..M.K.S. ...AS..... R......... ..V.S..... T...P---TE SSENK.KF.. I.AA...H.Y  644
Aca1A         SF..DT...G LMKKAKT.P. A.V.L..... .Y..M.K... ...AS..... R......... ..TQS.K..K ..D.P---SE .LEKKI...K A.ATH..L.Y  644
Xtr1A         SF..NS...G L.KKSRT.P. A.V.LS.... .Y..KEK... ...AS..... R......... ..I.S.D..L ..S.P---SQ .NEKR.Q..K E.AE...QMY  645
Lch1A         SF..DS...G L.KK.RT.P. A...L..... .YW.K.K.Y. ...AS..... R......... ..V.S....L S.V.P---KQ .NE.R.K..K I.SE...NMY  643
Ler1A         SF..DK...G V.KK.WI.P. A.V.M..... .Y..K.K... ...AS..... R......... ..SQSTQ..L ..VNS---SQ .NEEK.K... L.A....SMY  641
Xtr1C         TF..KE...G L.KK.RT.P. A......... .Y..K.K... ...AS..... R......... ..V.SSQ..K ..MNP---SQ .NEER.K... L.AE...M.Y  643
Ler1C         AF..TR...G V.KK.RT.P. A......... NY..K.T... ....S..... R......... ..V.T....T .LDG.---KH SNEE..N... I.AKN..HIS  644
Cmi1C         TF..TK...G L.KK.RT.P. A...L..... N...K.K... ....S..... R......... ..IQSS.... ....H----L SASE..K..H L.AKT..HIS  643
Aca1C         AFT.KD...G L.KK.RT.P. G...L..... ....KKK... ...AS..... R......... ..I.S....K ..M.P---TQ D.SAR.H... ..AE...N.Y  643
Lch1C         .FT.KD...G L.KK.RT.P. A......... .Y..K.K... ...AS..... R......... ..S.S.A..K ..V.A---GY .NAAR.K..K ..AE...QMY  642
Tni1A         II...E...G K.KK.RT.P. A......... .Y..K.K... ...AS..... R......... ..V.S.A... S.IRD----E .TEER.R.LK K.AE...N.Y  644
Tru1A         II...E...G K.KK.RT.P. A......... .Y..K.K... ...AS..... R......... ..M.S.A... S.IRD----E .TEER.R.LK K.AE...NMY  646
Oni1A         II..TD...G L.KK.RT.P. A......... .Y..K.K... ...AS..M. R......... ..I.T.A... ..VGD----E .REER.R.LK L.AE...N.Y  642
Gac1A         IM...D...G L.KK.RT.P. A......... .Y..KKK... ...AS..... R......... ..S.S.A... ..IKG----E .RDE..R.LK Q.AE...NMY  644
Dre1Aa        IF..ND...G L.KK.KT.P. G...L..... ....K.K... ...AS..... R......... ..N.T.A..H ..M.E---KA .REER.K.LK A.TE...N.Y  644
```

```
Tfu1A2     IF..NN...G L.KK.KT.P. G......... ....KKK... ...AS..... R......... ..M.T.D... ..M.E---KQ .REEK.R.LK L.AE...E.Y  644
Dre1Ab     II...D...G L.KK.RT.P. A...L..... .Y..K.K... ...AS..... R......... ..T.S.A... ..NSN----H .RE.K.Q.LK N.AE...QMY  648
Tfu1A1     II..QE...G L.KK.KI.P. A......... N.K.K.K... ...AS..... R......... ..I.TSA..K ..VNN----E .REKK.S.L. H.AGN..HMY  644
Ola1Ca     .LS.RD...G Q.KK.KV.P. A...MT.... YY.E..T... ...AS..... R......... ..N.SSA... .L.NGSPPSQ PADV.RR... G.SE...Q.Y  647
Oni1Cb     .FS.QE...G KVKK.RV.P. G...M..... Y..E..T... ...AS..... R......... ..N.SSA.I. .L.GG----E ATDV.KC... ..SE...L.Y  644
Ame1C      .FA.RD...G S.KKIKM.P. G...L..... FY....M... ...AS..... R......... .SS.S.A.IK .F.NG----E DVEV.RR.I. N.SE...M.Y  643
Gac1C      .FS.EE...G K.KK.RV.P. A...LT.... .Y.NQ.R... ...AS..... K......... ..N.S.A... .L.GG----E DADV.RR..H E.SE...Q.Y  646
Tru1C      ..V.RD...A KVKK.RVNP. A......... YY..QKR.S. ....S..... R......... ..N.S.A.I. .L.GG----E .TDV.RR... T.SE...R.C  645
Tni1C      ..V.RD...A KVKK.RVNP. A......... .Y..QKR.S. .......... R......... ..N.S.A.I. .L.GG----E .TDV.RR... T.CE...R.C  645
Ola1Cb     .F..RD...G K.KKLRV.P. A....S.... Y.....S... ...AS..... R......... .SNQS.A..K .L.GG----E GAD..KR... Q.SE...N.Y  645
Oni1Ca     .F..RE...G R.KK.RI.P. A....S.... YY...SG... ...AS..... R......... .SN.S.A.IK .L.SG----E .EEE.RR..Q L.SE...N.Y  643
Gac1C      .F..RD...G R.KKLRV.P. A....G.... Y.....G... ...AS..... R......... .SN.SSA... .L.T.----E AVD..RH... L.SER..N.Y  648
Tri1C      I.H.RD...G R.KKLRV.P. A.V..S.... YY....S... ...AS..... R......... ..N.SAA..K .L.NG----E DEES.RR... L.SET..N.Y  650
Dre1C      .F..RK...G L.KKMK..P. ..V.L..... YY....T... ...AS..... R......... .SN.S.A..L .L.GG----E DRE..RK.L. KTAE...N.Y  643
Dre1B      GCL.NE...G L.KK.RT.P. A...L..... QY..K.E... ....S...M. R......... ..C.STA... ....D---TT .NE.R....K Q.AE...NMY  642
Oni1B      GYL.AE...G L.KK.RT.P. A...L..... Q...QRV... ....S...M. RD........ ..S..VA... ....A---GA .NA.R..... K.AE...NMY  642
Gac1B      GYL.HE...G L.KK.RC.P. A...L..... Q...Q.V... ....S...M. RD........ ..T..VA... ....E---GA .KA.R..... K.AE...NMY  642
Ola1B      GHL.TE...G L.KK.RT.P. A...L..... Q...Q.V... ....S...M. RD........ ..S..VA... ....V---SA .NA.R.S..Q K.AE...NMY  642
Tru1B      GCL.HE...G L.KK.RT.P. A...M..... Q...Q.L... ....S...M. RD........ ..S..VA..K ...SA---DS .NA.R..... K.A....NMY  643
Tni1B      GCL.HE...G L.KK.RT.P. A...M..... Q...Q.L... ....S...M. RD........ ..S..VE... ...SA---DA .NA.R..... K.A....NMY  644
Hsa1B      CFQ.LP...G L.KK.RT.P. A.V....... ......K... ...AS...M. R......... ..S.STA..Q ..MEG---SH .KADLRD..Q K.AK...NMY  644
Mmu1B      CFQ.LP...G L.KK.RT.P. A.V....... ....K.K... ...AS...M. R......... ..N.SAA..Q ..MKG---SH KKQDLQD... K.SE...NMY  644
Mdo1B      CFH.AT...G L.KK.RT.P. A.V....... ....K.K... ...AS...M. RD........ ..T..TA..Q ..M.S---GY MK.DLQD... K.AE...L.Y  644
Sha1B      CFQ.LP...G L.KK.RS.P. A.V....... ....K.N... ...AS...M. RD........ ..A.TTA... ..T.S---GY MK.DLQD... K.AE...H.Y  644
Aca1B      SFR.VE...G R.KK.RT.P. .......... ....K.R... ...AS..... R......... ..S..TA... S.A.P---SC SSSER.E..H ..AE...H.Y  643
Fpe1B      CFQ..E...G L.KK.RT.P. A....S.... ....K.C... ...AS..... R......... ..A.STA... S.G.A---R. .VTERQR..K L.A....HMY  642
Xtr1B      CLC..D...G L.KK.RS.P. A.F....... .Y.EK.H... ...AS..... RD........ ..TQTSD..K ....P---TQ SQEKR...Y. A.AEH..LMY  642
LchB       CF..MS...G M.K.FKT.P. A......... ....K.A... ....S..... RD........ ..N.T.A..Q ..L.T---H. ..KER.S... K.SE...Y.Y  599
Dm         ILVHQDY..G .MKK.RI.P. AY..M..... YY..A.R.S. ...AS..... R........P ..I.SSAW.K ..QNP---NT .NDERVKMMQ A.C.R..LGY  641
ci         .SA.KH...G LVKKFKM.P. A...A...I. .L..K.R.S. ...AS..... R......... ..S.M.A.AK S...D---SF .NKDRY..LK K...R.INGY  636


Hsa1C      KAAMSGQGVD RHLFALYIVS RFLHLQSPFL TQ-------- ---------- ---------- --VHSEQWQL STSQIPVQQM HLFDVHNYPD YVSSGGGFGP  711
SSc1C      .......... .......... Q..R...... D.-------- ---------- ---------- --........ A......... ....I..... ..........  718
Mmu1C      .......I.. .......M.. .L..M..... ..-------- ---------- ---------- --.Q.Q..L. ....V...T. ..I....... ..........  709
Hsa1A      RL..T.S.I. ....C..V.. KY.AVE.... KE-------- ---------- ---------- --.L..P.R. ....T.Q..V E...LE.N.E ..........  714
Mmu1A      RL..T.A.I. ....C..V.. KY.AVD.... KE-------- ---------- ---------- --.L..P.R. ....T.Q..V E...FEK... ...C......  714
Mdo1A      RL..T.A.I. ....C..V.. KY.AVD.... KE-------- ---------- ---------- --.L.DP.R. ....T.Q..V E..NLERN.E ..........  714
Gga1A      RL..T.A.I. ....C..V.. KY.SVD.... KE-------- ---------- ---------- --.L..P.R. ....T.Q.HI ---.LKKN.E ML........  711
Fpe1A      RL..T.A.I. ....C..V.. KY.AVD.... KE-------- ---------- ---------- --.L..P.R. ....T.Q.HI ---.LKKN.E ML.C......  711
Aca1A      RL..T.N.I. ....C..V.. KY.AVE.... KE-------- ---------- ---------- --.L..P.R. ....T.Q.HI ---.LNKN.G ME........  711
Xtr1A      RL..T.S.I. ....C..V.. KY.GVD.... KE-------- ---------- ---------- --.L..P.R. ....T.Q..V ...QLEKF.E N.........  715
Lch1A      RL..T.A.I. ....C..V.. KY.GVD.... KE-------- ---------- ---------- --.L..P.R. ....T.Q..V E...LN.K.E ..........  713
Ler1A      RH..T.E.I. ....C..V.. KY.GMD.... KE-------- ---------- ---------- --.L..P.K. ....T..... N.H.----.K ..........  707
Xtr1C      RH..T.G.I. ....C..V.. KY.GVD.... KE-------- ---------- ---------- --.L..P.R. ....T.I..V E...LV.H.E ...C......  713
Ler1C      RL....C.I. ....C..V.. KY.GVS.... QE-------- ---------- ---------- --.L..P.C. ....T.I..I E...LV.H.E .I.I......  714
Cmi1C      RLS.A.C.I. ....C..V.. KY.GVH.... QE-------- ---------- ---------- --.L..P.S. ....T.L..V E...LL.H.E .I.C......  713
Aca1C      RQ..T.A.I. ....C..V.. KY.G.D.... RE-------- ---------- ---------- --.L..P.R. ....T.I..L E...LQ.H.. ...C......  713
Lch1C      RL..T.A.I. ....C..V.. KY.GVD.... NE-------- ---------- ---------- --.L..P.R. ....T.I..A E...LV.H.E ...C......  712
Tni1A      RL..T...I. ....C..V.. KY.GED.A.. KE-------- ---------- ---------- --.L..P.R. ....T.L..L E...LAKH.E ..T.......  714
Tru1A      RL..T.E.I. ....C..V.. KY.GEE.A.. KE-------- ---------- ---------- --.L..P.R. ....T....V E...LVKH.E ..........  716
Oni1A      RL..T...I. ....C..V.. KY.GED.... KE-------- ---------- ---------- --.L..P.K. ....T.L..V E...LVRH.E ..........  712
Gac1A      RL..I.E.I. ....C..V.. KY.GED.... KE-------- ---------- ---------- --.L..P.R. ....T.L..V E...LVRH.E ...A......  714
Dre1Aa     .L..T.K.I. ....C..L.. KY.GED.... KE-------- ---------- ---------- --.L..P.R. ....T.L... E...LKKH.E ..T.......  714
Tfu1A2     RM..T.K.I. ..I.C..V.. KY.GDD.A.. KE-------- ---------- ---------- --.L..P.R. ....T.L..I E...LKKH.E ..T.......  714
Dre1Ab     RL..T.H.I. ....C..V.L KY.GQD.... KE-------- ---------- ---------- --.L..P.R. ....T.L..G E...LVKN.E ..T.......  718
Tfu1A1     QM..T.K.I. ....C..V.. QY.QQD.... KK-------- ---------- ---------- --.L..P.R. ....T.L..P E...LL.H.E ..........  714
Ola1Ca     RM..T.A.I. ..I.C..V.. KY.G.E.... KE-------- ---------- ---------- --.LA.P.R. ....TSI..V E...LA.H.E .I.C......  717
Oni1Cb     RL..T.A.I. ....C..V.. KY.GVE.... KE-------- ---------- ---------- --.L..P.R. ....T.I..V E...IE.H.E ...C......  714
Ame1C      RL..T.A.I. ....C..V.. KY.GIE.... KE-------- ---------- ---------- --AL..P.R. ....T.F..L E...FV.H.. .ITC......  713
Gac1C      RM..T.A.I. ....C..V.. KY.QVE.... KE-------- ---------- ---------- --.L..P.R. ....T.L-.V EM..LV.H.E ...C......  715
Tru1C      RM.ST.A.I. ....C..V.. KC.GVE.... KE-------- ---------- ---------- --LE..A.R. .S.H..Y.MI D...TV.H.E .LCY......  715
Tni1C      RM.ST.A.F. ....C..V.. KY.GVE.... KE-------- ---------- ---------- --.L..A... AS.HV.H.MI D...TV.H.E .LCY......  715
Ola1Cb     RM..T.A.I. ....C..V.. KY.GVD.... KE-------- ---------- ---------- --.L..P.R. ....T..... E...LK.H.. F..L......  715
Oni1Ca     RM..T.A.I. ....C..V.. KY.GVE.... KE-------- ---------- ---------- --.L..P.R. ....T..... N...LK.H.. FI.L......  713
Gac1C      RM..T.A.I. ....C..V.. KY.GVD.... KE-------- ---------- ---------- --.L..P.R. ....T..... E...LK.H.. FI.L......  718
Tri1C      RM..T.A.I. ....C..V.. KY.GVD.... KE-------- ---------- ---------- --.L..P.R. ....T.I..I E...LK.H.. ...L......  720
Dre1C      RL..T.S.I. ....C..V.. KY.GVE.... KE-------- ---------- ---------- --.L..P.R. ....T..... E...LV.H.E FI.L......  713
Dre1B      RL..T.A.I. ....C..... KVMGID.... K.-------- ---------- ---------- --.L..P.R. ....T.Q..L N.I.TQKF.K ...A......  712
Oni1B      RL..T.S.I. ....C..... KY.GVD.... .K-------- ---------- ---------- --.L..P.K. ....T.Q..L N.V.INKF.K ..GG......  712
Gac1B      RL..T.S.I. ....C..L.. KY.GVD.... KK-------- ---------- ---------- --.L..P.R. ....T.Q..L N.V.INKF.K ...A......  712
Ola1B      RL..T.S.I. ....C..... KY.GAD.... NK-------- ---------- ---------- --.L..P.R. ....T.Q..L N.V.INKF.K ..GA......  712
Tru1B      RL..T.S.I. ...LC...I. KY.NVD.... KKVRTTPELF PFFARFLYIF VKRLCCFLCS LQ.L..P.R. ....T.Q..L N.V.IKKF.K ..GA......  743
Tni1B      RL..T.S.I. ...LC..... KY.GVD.L.. KK-------- ---------- ---------- --.L..P.R. ....T.Q..L N.V.ISKF.K ..GA......  714
Hsa1B      RL..T.A.I. ....C..L.. KY.GVS.... AE-------- ---------- ---------- --.L..P.R. ......QS.I RM..PEQH.N HLGA......  714
Mmu1B      RL..T.A.I. ....C..... KY.GVS.... AE-------- ---------- ---------- --.L..P.S. ......QF.I CM..PKQ..N HLGA......  714
Mdo1B      RL..T.A.I. ....C..V.. KY.G.H.... A.-------- ---------- ---------- --.L..P.R. .....AQF.I .M..PEK..N HIAA......  714
Sha1B      RL..T.A.I. ....C..V.. KY.G.H.... A.-------- ---------- ---------- --.L..P.R. ....TAQF.I RM..PEK..N HLAA......  714
Aca1B      RL..T.S.L. ....C..V.. .Y.GVE.... DK-------- ---------- ---------- --.L..P.R. ....T.Q..I KM..LEAH.E CA........  713
Fpe1B      RL..T.A.I. ....C..V.. .Y.GI..... A.-------- ---------- ---------- --.L..P.R. ....T.Q..L KM..LNK... H.........  712
Xtr1B      RW..T.K.I. ....C..... KY.GTD.A.. QK-------- ---------- ---------- --.L..P.R. ....T.Q..L K...LDKF.. H..A......  712
LchB       RL..T.A.I. ....C..V.. K..GVH.... N.-------- ---------- ---------- --.LA.P.A. ....T.Q..T S..NLSK... ..........  669
Dm         QD..C.R.I. ....C..V.. KY.EVD.... NE-------- ---------- ---------- --.L..P.R. ....T.HG.T PKM.LKKH.N CI.A......  711
ci         .E..T...I. ..I.C..V.. KY.K.E.... QK-------- ---------- ---------- --.LQ.P.R. ....T.H..A MNV.LNKH.N FL.G......  706


Hsa1C      ---------- ---------- ---------- ----ADDHGY GVSYIFMGDG MITFHISSKK SSTKTDSHRL GQHIEDALLD VASLFQAGQH FKRRFRGS--  775
```

```
SSc1C      ---------- ---------- ---------- ----..E... ......T.ED T......... ...R...... ....R..... ..A...E... L...A-----  779
Mmu1C      ---------- ---------- ---------- ----.H.... .I......EN A......... ...E...... .....N.... .....RV... ...Q...---  772
Hsa1A      ---------- ---------- ---------- ----VA.D.. .....LV.EN L.N......F .CPE.----- ---------- --GIISQ.PS SDT-------  756
Mmu1A      ---------- ---------- ---------- ----VA.D.. .....IV.EN F.H......F ..PE.....F .K.LRQ.MM. IIT..GLTAN S.K-------  773
Mdo1A      ---------- ---------- ---------- ----VA.D.. .....IV.EN L.N..V...F ..PE.....F .N.LKQ.MI. IIT..GLQTN DQKH------  774
Gga1A      ---------- ---------- ---------- ----VA.D.. .....ILDEN S.H..V...F .CSE.....F .KN.QK..V. IMG...PTKN CTK-------  770
Fpe1A      ---------- ---------- ---------- ----VA.D.. .....ILDEN S.H..V...I .CSE.....F .KN.QK.MV. IMG..NLSKN CTK-------  770
Aca1A      ---------- ---------- ---------- ----VA.D.. .....IV..N L.N..V...Y .CPE.N...F .KN.KR..C. IRDM.GI.KN STK-------  770
Xtr1A      ---------- ---------- ---------- ----VA.D.. .....IV.EN L.N......F ..PE.....F .K..KQ.MI. ILA..NISTN NSNKGKK---  778
Lch1A      ---------- ---------- ---------- ----VA.D.. .....LV.EN L.NL.V...L ..LE.....F .K..RQ.MQ. ILA..NLNNK SSK-------  772
Ler1A      ---------- ---------- ---------- ----VA.D.. .....MV.EN L.NM.....F ..PE.....F .NYFQQ.M.. ILA..DLDKR TSK-------  766
Xtr1C      ---------- ---------- ---------- ----VA.D.. .....IV.EN L.N......F ..HE.....F .K..QG..R. ILA..KLH-- ----------  767
Ler1C      ---------- ---------- ---------- ----VA.D.. .....IV..H L.N....C.V ..PL..A..F .EN.SQ.M.. IIT..NLDGK KA--------  772
Cmi1C      ---------- ---------- ---------- ----VA.D.. ....FIV.EN L.N....C.V ..QY..A..F .K..VE..M. ILA..KLDEK SHK-------  772
Aca1C      ---------- ---------- ---------- ----V..N.. .....IV.ED L.N..V.C.V ..PE..A..F .LN.RN.MT. I.L..K.NKK .AA.------  773
Lch1C      ---------- ---------- ---------- ----VA.D.. .....IV.ED L.N....G.I .GLD..T..F .H..RK.... IL...AVAKK N---------  769
Tni1A      ---------- ---------- ---------- ----VA.D.. .....IL.EN L.N......R ..PE.....F .TN.KR.MV. MLN..ELDRK T.--------  772
Tru1A      VGNPFICIMV SNI------- ------LKFL SSLKVA.D.. .....IV.EN L.N......R ..PE.....F .TN.KQ.MV. MLN...LEKK A.--------  795
Oni1A      ---------- ---------- ---------- ----VA.D.. .....IL.EN H.N......R ..PE.....F .TN.RQ.M.. ILG..ELNNK AAK-------  771
Gac1A      ---------- ---------- ---------- ----VA.D.. .....IL.EN L.N......H ..PE.....F SG..RQ.M.. IKG...LEKK TQ--------  772
Dre1Aa     ---------- ---------- ---------- ----VA.D.. .....IL.ED L.N......H ..HE.....F .SN.KQ.M.. ILD...LDIK PNKK------  774
Tfu1A2     ---------- ---------- ---------- ----VA.D.. .....IL.ED L.N......Y ..IE.....F .N..KQ.M.. ILA..ELDKK TVK-------  773
Dre1Ab     ---------- ---------- ---------- ----VA.D.. ..A.VIV.EK L.N......R ..PE.....F .NN.RR.MI. MLD...LDKK DPK-------  777
Tfu1A1     ---------- ---------- ---------- ----VA.D.. ..A..IV.EN L.N....C.Y ..PEA.AQ.F .NN.KQ.M.. MLE...LETK ALK-------  773
Ola1Ca     ---------- ---------- ---------- ----VA.D.. ....N.I.EN V.N....C.H .CPS..A.KF .DQ.RK.VN. LLQ.LSPN.K REWN------  777
Oni1Cb     ---------- ---------- ---------- ----VA.D.. ....CVL.EN ..N....C.H .CPD..A.KF .TQ.RQ..R. LLK.LSPN.T EPS-------  773
Ame1C      ---------- ---------- ---------- ----VA.D.. ....SLI.EK I.S..V...H .CPD....KF .TQ.RK..H. LLQ.MTIN.S DAAGSEQ---  776
Gac1C      ---------- ---------- ---------- ----VA.D.. ....CAL.EK .LS...TC.H .CPN..ANTF .GE.RK..R. LLQ.LSPE.A EP.KAAA---  778
Tru1C      VSASLMPLPQ HC-------- ---------L CVFQVT.D.. ..C.LML.GD VL.L...C.N .CPD..ARKF .AR.RA..H. LMQ.LSPNPK EP--------  790
Tni1C      ---------- ---------- ---------- ----VT.G.. ..C.LML.GD VL.L.V.C.N .CPA..ARKF .AQ.RT..H. LIQ.LSPNPK EPLK------  775
Ola1Cb     ---------- ---------- ---------- ----VA.D.. .....IV.ED ..N..V...H .CGE.....F .AQ.AK..Q. IMAVLS.DTA SSSKG--P--  777
Oni1Ca     ---------- ---------- ---------- ----VA.D.. .....IV.ED .VN..V...Y .CSQ.....F .IQ.SK.MQ. IMA.LS.DPK TSSSSK.---  776
Gac1C      ---------- ---------- ---------- ----VA.D.. .....IV.ED ..N..V...H .CSE.....F .VQ.SK..Q. IMN.LA---- ----------  770
Tri1C      VSMKYLSLCV HYVCFLQVAF LVSVMLFYDL CSYKVA.D.. .....IT.ED ..N..V...H .CNQ.....F .AQ.SK..K. IMTVLSSSPK VQ--------  812
Dre1C      ---------- ---------- ---------- ----VA.D.. .....I..ED ..N..V.C.H .CKE....KF .CQVSQ.MV. LMT.LNPDFR DTTEANS---  776
Dre1B      ---------- ---------- ---------- ----VA.D.. .....IV.EN L........F ..PE...F.F ..N.RQ.MQ. IRA..NQKEK KM--------  770
Oni1B      ---------- ---------- ---------- ----VA.D.. .....IV.EN L........F ..PD...Y.F ....RK.M.. IQA..KPEND KTAQNAKY--  776
Gac1B      ---------- ---------- ---------- ----VA.D.. .....IV.EN L........F ..PD.N.C.F ....C..M.. IQA..KPEND KTTVE-----  773
Ola1B      ---------- ---------- ---------- ----VA.D.. .....IV.EN L........F ..PN...Y.F ..N.QR.MI. IQT..TLEHD K.TPD-----  773
Tru1B      ---------- ---------- ---------- ----VA.D.. .....IV.EN L........F .CPH.VQP.E RN.------P FT.V.CYRSN LSLPSCRTRT  803
Tni1B      ---------- ---------- ---------- ----VA.D.. .....II.EN L........F .CPH.....F ....RK.M.. IQA..K.DD. TRTME-----  775
Hsa1B      ---------- ---------- ---------- ----VA.D.. ....MIA.EN T.F......F ..SE.NAQ.F .N..RK.... I.D...VPKA YS--------  772
Mmu1B      ---------- ---------- ---------- ----VA.D.. ....MIA.EN TMF......Y ..SE.NAQ.F .N..RQ.... I.E..KISKT DS--------  772
Mdo1B      ---------- ---------- ---------- ----VA.D.. ....MIA.EN T.F..V...F ..SE.NAQ.F .N..RQ..Q. I.A..EISVP KTES------  774
Sha1B      ---------- ---------- ---------- ----VA.D.. ....MIA.EN T.F......F ..SE.NAQ.F .N..HQ..E. I.A..ELPSP KPES------  774
Aca1B      ---------- ---------- ---------- ----VA.D.. .....IA.EN LV...V...F ..PE...K.F .RN.HR.M.. I.A..D.SAK RMS-------  772
Fpe1B      ---------- ---------- ---------- ----VA.D.. .....IA.EN L....V...F ..SE.GLITR .P-------- ---------- ----------  750
Xtr1B      ---------- ---------- ---------- ----VA.D.. ......A.EN L..L.....F ..PE.N...F ..R.CQSMR. L.Q.LSPPVT VRP-------  771
LchB       ---------- ---------- ---------- ----VAED.. .....IV.ED L........Y ..PE.....F AKN.QQ.MM. LKT.QESSRV KPKN------  729
Dm         ---------- ---------- ---------- ----VA.D.. .....IA.EN L.F....A.T TCQQ..V..F A.N.SQ..S. IR.M.EQHMK DHPKPAK---  774
ci         ---------- ---------- ---------- ----VA.D.. .....ICHEN L.M..V...Y ..SE...D.F AGN..K.M.. LRN.CESIIL YS--------  764
```

```
Hsa1C      ---------- ---------- --GK--ENSR HRCGFLSRQT GASKASMTST D-F 803
SSc1C      ---------- ---------- ---------- ---------- ---------- --- 779
Mmu1C      ---------- ---------- ------...D Y.YN...CK. VDPNTPTS.. N-L 798
Hsa1A      ---------- ---------- ---------- ---------- ---------- --- 756
Mmu1A      ---------- ---------- ---------- ---------- ---------- --- 773
Mdo1A      ---------- ---------- ---------- ---------- ---------- --- 774
Gga1A      ---------- ---------- ---------- ---------- ---------- --- 770
Fpe1A      ---------- ---------- ---------- ---------- ---------- --- 770
Aca1A      ---------- ---------- ---------- ---------- ---------- --- 770
Xtr1A      ---------- ---------- ---------- ---------- ---------- --- 778
Lch1A      ---------- ---------- ---------- ---------- ---------- --- 772
Ler1A      ---------- ---------- ---------- ---------- ---------- --- 766
Xtr1C      ---------- ---------- ---------- ---------- ---------- --- 767
Ler1C      ---------- ---------- ---------- ---------- ---------- --- 772
Cmi1C      ---------- ---------- ---------- ---------- ---------- --- 772
Aca1C      ---------- ---------- ---------- ---------- ---------- --- 773
Lch1C      ---------- ---------- ---------- ---------- ---------- --- 769
Tni1A      ---------- ---------- ---------- ---------- ---------- --- 772
Tru1A      ---------- ---------- ---------- ---------- ---------- --- 795
Oni1A      ---------- ---------- ---------- ---------- ---------- --- 771
Gac1A      ---------- ---------- ---------- ---------- ---------- --- 772
Dre1Aa     ---------- ---------- ---------- ---------- ---------- --- 774
Tfu1A2     ---------- ---------- ---------- ---------- ---------- --- 773
Dre1Ab     ---------- ---------- ---------- ---------- ---------- --- 777
Tfu1A1     ---------- ---------- ---------- ---------- ---------- --- 773
Ola1Ca     ---------- ---------- ---------- ---------- ---------- --- 777
Oni1Cb     ---------- ---------- ---------- ---------- ---------- --- 773
Ame1C      ---------- ---------- ------NH.S KTQNKKE--- ---------- --L 788
Gac1C      ---------- ---------- ------QRPE LKKD------ ---------- --- 786
Tru1C      ---------- ---------- ---------- ---------- ---------- --- 790
Tni1C      ---------- ---------- ---------- ---------- ---------- --- 775
Ola1Cb     ---------- ---------- ---------- ---------- ---------- --- 777
Oni1Ca     ---------- ---------- ------RGH- ---------- ---------- --- 779
Gac1C      ---------- ---------- ---------- ---------- ---------- --- 770
Tri1C      ---------- ---------- ---------- ---------- ---------- --- 812
```

```
Dre1C       ---------- ---------- ------DEQT ---------- ---------- --- 780
Dre1B       ---------- ---------- ---------- ---------- ---------- --- 770
Oni1B       ---------- ---------- -----VHLEN GKKHI----- ---------- --- 786
Gac1B       ---------- ---------- ------QLGN GKKHM----- ---------- --- 782
Ola1B       ---------- ---------- ------CLQ- ---------- ---------- --- 776
Tru1B       ASASTFRRPC WTSRRSSQKR SK.RRWRTEN T.RWKME.S. YERYCGWNRQ KYG 856
Tni1B       ---------- ---------- ---------N GKHA------ ---------- --- 780
Hsa1B       ---------- ---------- ---------- ---------- ---------- --- 772
Mmu1B       ---------- ---------- ---------- ---------- ---------- --- 772
Mdo1B       ---------- ---------- ---------- ---------- ---------- --- 774
Sha1B       ---------- ---------- ---------- ---------- ---------- --- 774
Aca1B       ---------- ---------- ---------- ---------- ---------- --- 772
Fpe1B       ---------- ---------- ---------- ---------- ---------- --- 750
Xtr1B       ---------- ---------- ---------- ---------- ---------- --- 771
LchB        ---------- ---------- ---------- ---------- ---------- --- 729
Dm          ---------- ---------- ------SLTN GKST------ ---------- --- 782
ci          ---------- ---------- ---------- ---------- ---------- --- 764
```

**Supporting information 1 Table:** List of sequences used for the molecular phylogenetic analysis and respective accession numbers (GenBank or Ensembl).

| Species | Gene | Accession number |
|---|---|---|
| *A. carolinensis* | CPT1A | XP_003214835.1 |
| | CPTC | XP_003222743.1 |
| | CPT1B | ENSACAP00000001055 |
| *C. intestinalis* | CPT1 | ENSCINP00000007072 |
| *C. milii* | CPT1C | AGD98733.1 |
| *D. melanogaster* | CPT1 | CAB52415.1 |
| *D. rerio* | CPT1Aa | XP_002666893.2 |
| | CPT1Ab | XP_005166530.1 |
| | CPT1Ca | XP_005164116.1 |
| | CPT1Cb | XP_002666747.2 |
| | CPT1B | XP_005159068.1 |
| *F. peregrinus* | CPT1A | XP_005236351.1 |
| | CPT1B | XP_005243516.1 |
| *G. aculeatus* | CPT1A | ENSGACP00000014767 |
| | CPT1Ca | ENSGACP00000010584 |
| | CPT1Cb | ENSGACP00000008742 |
| | CPT1B | ENSGACP00000016316 |
| *G. gallus* | CPT1A | NP_001012916.1 |
| *H. sapiens* | CPT1A | NP_001867.2 |
| | CPT1B | NP_689451.1 |
| | CPT1C | NP_001186681.1 |
| *L. chalumnae* | CPT1A | ENSLACP00000021346 |
| | CPT1C | ENSLACP00000014661 |
| | CPT1B | ENSLACP00000008053 |
| *L. erinacea* | CPT1A | KF570112 |
| | CPT1C | KF570111 |

| | | |
|---|---|---|
| *M. domestica* | CPT1A | XP_001363149.1 |
| | CPT1B | XP_001366412.1 |
| *M. musculus* | CPT1A | NP_038523.2 |
| | CPT1B | NP_034078.2 |
| | CPT1C | NP_710146.1 |
| *O. latipes* | CPT1Ca | XP_004080367.1 |
| | CPT1Cb | XP_004071913.1 |
| | CPT1B | XP_005470933.1 |
| *O. niloticus* | CPT1A | XP_003440402.1 |
| | CPT1Ca | XP_003438524.2 |
| | CPT1Cb | XP_003446513.1 |
| | CPT1B | XP_005470933.1 |
| *T. nigroviridis* | CPT1A | CAG01138.1 |
| | CPT1C | CAG07569.1 |
| | CPT1B | CAG11364.1 |
| *X. tropicalis* | CPT1A | XP_004913492.1 |
| | CPT1C | NP_001107300.1 |
| | CPT1B | NP_001072766.1 |
| *S. scrofa* | CPT1C | XP_005664799.1 |
| *T. rubripes* | CPT1A | XP_003967444.1 |
| | CPT1Ca | XP_003961459.1 |
| | CPT1Cb | XP_003964428.1 |
| | CPT1B | XP_003967147.1 |
| *T. fulvidraco* | CPT1Aa | AFO11024.1 |
| | CPT1Ab | AFO11025.1 |

**Supporting information 2 Fig:** MAFFT Sequence alignment without gaps

```
Hsa1C    LRSWKRSRFW NDFLTGVFPA SPLSWLFLFS AIQLAWFLQL DPSLGLMEKI KELLPDWRLA AALFASCLWG ALIFLLEPHG AMSSPTKTWL ALVRIFSGRH  100
SSc1C    .......... ....I..... .......... .....C.... .......... .......... .......... .......T.. V......... ..........  100
Mmu1C    ......W.V. ....A..V.. T......... T....CL... .......... .......Q.S ..V....... .......... .......... ..........  100
Hsa1A    .H...KI..K .GII...Y.. ..S...IVVV GVMTTMYAKI .....IIA.. NRT.ETAKVS GV..GTG..V ...VMFTE.. K..RA..I.M GM.K.....K  100
Mmu1A    .H...KI..K .GII...... ..S...IVVV GVISSMHTKV .....MIA.. NRT.DTTKVS GV..GTG..V .I.MMFAE.. K..RS.RI.M .M.KV....K  100
Mdo1A    .H...KI..K .GII...Y.. ..S...IVVV GVMSTMYAKV .....IIA.. NQT.DMTKVS GI..GTG..V T..VMFAEY. KL.RG.RI.M GM.K.....K  100
Gga1A    VH...KI..K .GII...Y.. ..S...IVVV GVMSTMYAKI .....IIA.. NRT.DTTQVS GI..GTG..V ...VMFAE.. KL.AG..L.M T..KL....K  100
Fpe1A    VH...KI..K .GII...Y.. ..S...IVVV GVMSTMYAKI .....IIA.. NRT.DTTQVS GV..GTG..V ...VMFAE.. KL.AG..F.M ...KL....K  100
Aca1A    .Y...KI..K .GII...Y.. ..S...IVVV GVMSTMYAKI ......IA.. NRT.DITQVS GV..GTG..V ....MFAE.. KL.TG..I.M T..KL....K  100
Xtr1A    .H...KI..K .GI......S ..S...IVVV GVTSSMYTKV ...F.II... NNA.SATQVS GV..GTG..V S..AMF.E.. KL.AS.RI.M GM.KLL...K  100
Lch1A    .H...KI..K .GII...Y.. ..S....VVV GVVSTMYTRV .....MIAR. SQR..VTQVS GV..GTG..V ..F.MFIE.. KVPFG.RI.I R..KL..VCK  100
Ler1A    .....KR.YK .GI....Y.. ..S....VVV .VIATMYARV .....MI... RCH..TNE.S GL..STG..I ..V.MFV... KTPTSV.L.M LI.K.....K  100
Xtr1C    .....KA.VK .S.I...Y.. ..S....VVI ..MGTLYARV ...M.MI... ..H..ASQVS .L..STG..F ...MMY.Q.. K..AT..L.. ...K.....N  100
Ler1C    I....KN.IK .N.M...Y.. ..S..I.VVL TVVGTMYTRV ...M..IAML R.HI.LRQVS .M...TG..L ....MF.E.. KV..T..V.F T..K.....K  100
Cmi1C    V.A..KA.IR .NIV...Y.. ..S....VAL TVVGTMYTRV ...M.MIG.. ..H..VKQMS .L...TV..L ...YMF.E.. KA.NM.RI.F TM.K.....K  100
Aca1C    V....KNQLR .G.V...Y.. ..S....MVT ..LVTQYSR. ...M.MID.. ..H..VSM.S ..T.STV..L ...MMY.E.. K..NT..I.. ...KM.A..K  100
Lch1C    Q....KA.LK .S.I..MY.. ..S.....VI ..MATLYARV ...M.MID.. ..H..VKI.C TN.YQCGIYS EILLMF.Q.N RV.LT..I.I T..K.....K  100
Tni1A    .H...KI..K .GVM...Y.G ..AGFMIVVG SYMSNKYNR. ...M..VV.L GQYI.IGQVG GV.VGTS..V TI.MMH.R.. SV.WTSRA.M L..KV....K  100
Tru1A    .H...KI..K .GVM...Y.G ..GGFMIVVG SYMSNKYN.. ...M.FVV.L GQYI.ISQVG GV.VGTG..V TI.MMY.R.. SI.WTSRA.M L..KV....K  100
Oni1A    IH...KV..K .GIM...Y.G ..AGFSIVVG SYMANK.K.. ......FT.L GQHI.ISQVG GI.VGTS..V TI.MMQ.R.. SL.WSSRI.M V..KV....K  100
Gac1A    .H...KI..K .GVM...Y.G ..GGFMVVVV SYMSNKYQ.. ......IA.L GQHM.ISQVG GV.VGTG..V SI.MMYTS.. SVAWS.RL.M V..KV....K  100
Dre1Aa   IH...KI..K .GVM...Y.G ..SGLVVVLV GYMSTKYAKI .....ILT.L STH..VSQVG GV.VGTG..I .VT.MFNQ.. TL.LK..I.. V..KL...PK  100
Tfu1A2   .H...KI..K .GVM...Y.G ..TGL.GVVG GYLITKYANI ..T..VVA.L APH..VCQVG GI.VGTGV.T .V..MYNR.. T..IR..I.. ...KL...PK  100
Dre1Ab   IH...KI..K .GI.N..Y.G .APGFVLVLA GYLG.QY.KV ......LF.L GNYV.ISQVG GVMVGTG..I .I..MFAS.. R.TWKIRL.. VF.KV...QT  100
Tfu1A1   .H...KI..K .GIM...Y.G .APGLMLVLA GYMG.KYA.V ......VFR. GKYV.ISQVG GI.VGTG..V VI..MR.R.. SLTWK.QI.. V..KV...MK  100
Ola1Ca   V....KI.IK .SVI...Y.. ..S....VVI ..LATMYTRS ...M.IIA.. Q.H..VSQVS .V..STM..L L...MF.Q.. K..TT..I.V .........K  100
Oni1Cb   M....KV.VK .SVI...Y.. ..S....VVI ..LATMYTRS ...M..IA.. Q.H..VSQVS .V..STM..L M...MF.Q.. K..TT..I.V .........K  100
Ame1C    .....KV.LK .NVI...Y.. ..S....VVI ..LATMYTRS ...M..IA.. Q.H..ASQVS .V..STL..L S...MF.Q.. NV.TT..V.V M........Q  100
Gac1C    V....KIGLK .SVV...Y.. .LS....VVI ..LATMYTRS ...M..IA.. Q.H..RPY.S .S.AVTM..L L...MF.K.. K..NT..V.V VS.ASHPS.K  100
Tru1C    VN...KI.IK .SVIR..Y.. ..F....VVI ..LATMYTRS ...M..IA.. Q.N..VSK.S .VI.STM..L L...MF.... K..TT..V.V .........K  100
Tni1C    VH...KI.IK .SVIK..Y.. ..F....VVI ..LATMYTRS ...M..IA.. Q.N..VSK.S .VI.STM..L L...MF.... K..TT..V.V .........K  100
Ola1Cb   V....K..MR .RVIK..Y.. ..S....VAI G.LATIYM.S ...M..IAE. QQR..LSQVS .LV.STL..L S..LMF.Q.. RV.NT..V.V T.L.LL...K  100
Oni1Ca   V....K..VR .SLIK..Y.. ..S....VAI G.LATMYM.S ...M..IT.. QQH..LSQVS .LV.STL..L S..LMF.K.. RI.NT..V.V P...LL.S.K  100
Gac_0    V....K..MR .RVIK..Y.. ..S....VVI ..LATMYMRS ...M..IKM. QQH..LRQ.S .LV.STL..L S..LMF.Q.. RV.NT..V.V T.L.LL.S.K  100
Tri1C    V....K..VR .SLIK..Y.. ..S....VSI ..LATMYMRS ...M..IT.. QHH..LSQ.S .LV.STL..L S..LMF.Q.. RV.NF..V.V T.LGLL.S.K  100
Dre1C    .....K..IK .RIIK.AY.. ..S....IVI ..LATMYM.S ...M..IA.. Q.H..LSQ.S .L..STL..M S..LMF.Q.. H..TK..V.A T..KLL.S.K  100
Dre1B    VT...KI..K .SV....Y.. ..S....VVI ..MSTMYARI ...M.TID.. .TS..VSQ.S .I...TG..L SV..IF.S.. K..YS..V.. S..KLL...R  100
Oni1B    VTA..KIQ.K .GV.A..Y.. ..S...IVVI .MMSSLYIHI .....M.DA. ..N..YR..S .I...TG..L F..YIF.S.. K..TS..V.. C..KM....R  100
Gac1B    VTA..KIQ.K .GV.A..Y.. ..S...IVVI .MMSSLYIRI .....MI.AL Q.N..HR..S .I...TG..L F..YIF.S.. K..RS..V.. S..KM....R  100
Ola1B    .TV..KIQ.K .SV....Y.. ..S...IVVI .MMSSLYTNT .L...MIDA. ..N..HR..S .I..GTG..L F..YIF.S.. K...S..L.. Y..KM....R  100
Tru1B    .TA..KIQ.K .GV.A..Y.. ..S...IVVI .MMSSLYTRV .....MI.AM ..N..YR..S .I...TG..L F..YIF.S.. K..TS..V.. T..KM....R  100
Tni1B    .TA.RKIQ.K .GV.A..Y.. ..S...IVVI .MMSSLYI.V .L...MI.AM ..N..YR.VS .I...TG..L F..YIF.S.. K..TS..V.. T..KM....R  100
Hsa1B    IN...KI.IK .GI.R..Y.G ..T...VVIM .TVGSS.CNV .I....VSC. QRC..QG..S M.I.STGV.V TG..MF.M.. KT.NL.RI.A MCI.LL.S..  100
Mmu1B    IN...KI.IK .GI.R..Y.G ..T...VVVM .TVGSNYCKV .I.M..VDC. QRC..ERE.S MVI.STGV.A TG..MF.M.S KT.HA..I.A IC..LL.S.R  100
Mdo1B    IN...KI.IK .GI.R..Y.G ..T...VVVM TTMGSSYCNV .L.M.MICC. RKYI.EG.IS VGI.STGV.V TG..MF.L.. QT.RT..I.A IC..LL.N.R  100
Sha1B    IN...KI.IK .GI.R..Y.G ..T...VVVM .TMGSSYCNV .I.M.MICH. RKYI.EG.IS VGI.STGV.V TG..MF.M.. QT.RI..I.A IC..LL.N.R  100
Aca1B    VS...KA.LK .SI......G T.SG..AVVA VTVG.TYFGI .V.V..IFR. QKR..NS..S .LI.S.GA.M LGVLMF.... KTRPS..I.A SM..VM....  100
Fpe1B    IS...KV.AK .S.....Y.. ..S..MVVVM .TAGSFYC.V .....MIAR. RHC..ES.VS TVI.STGA.L SAVLMF.... K..RS.RI.V ..MKVL.I.K  100
Xtr1B    .VACRKITLK SSV.S..Y.. ..ST.FAVVA MTLGSLYGK. ......IT.. NSI..GK..S .VI.S.GV.V SG.LMF...K KT.MK..I.A GCMK.M.S.Q  100
LchB     .MN.LHVFVQ .SL....Y.. ..S...VVVV .TIGTRYVKM .....IIDY. RSAI.RS..S .T...TGV.V TG.LMF.... STRIR.RI.A T..K......  100
Dm       V....KA.AR .GVRN..Y.. HIQ.LWLISA IALGLH.AGY QAPFN.TNR. LVH..SNWT. CF.A.LVV.L SIC.MY.SRS RV.L..ML.V .V..VL.SNK  100
ci       V...RKT..R .RV.S....V KYT.LVG.TA IVL..S..SY .ITW.FKNNR TNII.PRNTS YLVSSTL..L LAVLMF..R. K..LK..L.A SYLQLLCYTQ  100

Hsa1C    PMLFSYQRSL PRQPVPSVQD TVRKYLESVR PILSDEDFDW TAVLAQEFLR LWYLRLKSWW ASNYVSDWWE EFVYLRSRNP LVNSNYYMMD FLYVTPTPLQ  200
SSc1C    .......... ......A... .......... .V.CE...E. ISA..R...K ....Q..Y.. .......... .........S .......... ..N.....V.  200
Mmu1C    .R...F..A. .......A.E .......... .V.G.DA..R ATA..ND... .L..Q....C T......... ........GS .I..T..... ..........  200
Hsa1A    ...Y.F.T.. ..L...A.K. ..NR..Q... .LMKE...KR MTA...D.AV G...K..... .T........ .YI...G.G. .......A.. L..IL..HI.  200
Mmu1A    ...Y.F.T.. ..L...A.K. ..SR...... .LMKEG..QR MTA...D.AV N...K..... .T........ .YI...G.G. I......A.E M..I...HI.  200
Mdo1A    ...Y.F.T.. ..L...A.K. ..NR.....Q .L.NKDN.QR MKG..ED.ST N...K..... .T........ .YI...G.S. I......A.. L..IL..TI.  200
Gga1A    ...Y.F.T.. ..L...A.K. ..NR...... .LMN..E.KR MEG..KD.AF N...K..... .T........ .YI...G.G. I.....FA.. ..HLS..TI.  200
Fpe1A    ...Y.F.T.. ..L...A.N. ..NR...... .LMD..E.RR MEG..KD.AF N...K..... .T........ .YI...G.G. I.....FA.. ...LS..T..  200
Aca1A    ...Y.F.T.. ..L.....KN ..NR.....H .LMNE.Q.KR MEA.GKD.AT N...K..... .A........ .YI...G.G. I.....FA.. ....F..SV.  200
Xtr1A    ...Y.F.T.. ..L...P.K. ..KR..D..K .LMDK.K.ER MEG..KD.AN N...K..... .T........ .YI...G.G. I......A.. .....L.HI.  200
Lch1A    ...Y.F.T.. ..L...A.K. .MTR...... .LMD..Q.RR MTK..KD.EL K...K..... .T........ .YI...G.G. I......A.. L.........  200
Ler1A    .LTY.F.T.. ..L...T.K. .M........ .L.N.KE.QR MQA..KD.EL K...K..... .T........ .YI...G.G. I......A.. Y..IV...V.  200
Xtr1C    ...Y...AV. ..L...G.KE ..QR..D... .LMN..EYKR MTG..KD.EV N...K..... .A........ .YI...G.G. I......A.. ........V.  200
Ler1C    ...Y.F.N.. ..L...P... .L.RF..... .LMN....RR .KA..KD.E. N...I..... .......... .Y...QG.E. I......A.. .......TS.  200
Cmi1C    ...Y.F.A.. ..L...TI.. .MQR...... .LMN..E.HR MEA..KD.EV K...K..... .T........ .Y....G.E. I......A.. ........I.  200
Aca1C    ...Y...A.. ..L...ALK. .MQ.....I. .LTT.AE.QR M.A..RD.EQ T...K..... .......... .Y....G.G. .......A.. .I......V.  200
Lch1C    ...Y...AC. ..L...PIAK .MQR.....H .LMD..K.KR MTA..KD.EV N...K..... .......... .YI...G.G. I......G.. .I.AS..YI.  200
Tni1A    ...Y.F.N.. ..L...DIS. .C.RH..... ALMD..Q.ER MTA.TKD.EK N...K..... .......... .YI...G.G. I......A.. ....I..SI.  200
Tru1A    ...Y.F.N.. ..L...A.S. .C.R.....H .LMDE.R.ER MTA..KD.EK N...K..... .......... .YI...G.S. I......A.. ....F..SI.  200
Oni1A    ...Y.F.N.. ..L....IK. .CER...... .LMD.QQ.ER MKG.T.D.EK N...K..... .......... .YI...G.G. I......A.. ....F..SI.  200
Gac1A    ...Y.F.N.. ..L....IK. .CKR...... .LME..QYQR MEG.TKD.EK N...K..... .......... .YI...G.G. I......A.. ....F..SI.  200
Dre1Aa   ...Y.F.S.. ..L...P.K. ...R....A. .LMD..QYKR MEG..KD.EK N...K..... T......... .YI...G.S. I......A.. ....N..S..  200
Tfu1A2   ...Y.F.S.. ..L...P.EH ..KR...... .LMD..QYKR MEA..KD.QS N...K..AF. .......... DY..V.G.G. I......A.. ....F..H..  200
Dre1Ab   ...Y.F.N.. .HLF....KE .T.R.....Q .L....EHQR MQR..LD.E. N...K..... .T........ .YI...G.G. I......V.. ...AF..NI.  200
Tfu1A1   .N.Y.F.T.. .NL.L...K. .MKR...... .L.D.TEYKK MEE..SD.QK T...K..... .T........ .YI....... I......A.. ..H.L..H..  200
Ola1Ca   .L.Y...G.. .NL...AIK. ..KR...... .LMN.GEYER MTK..T..ES S...K..AL. .......... .Y....G.S. I......G.. ........I.  200
Oni1Cb   .L.Y...G.. .NL...TIK. ..KR...... .LMD.KEYER MTK..A..ES S...K..AL. .......... .Y....G.G. I......G.. ........I.  200
Ame1C    .L.Y...G.. .NL...AIK. ..KR...... .LK..SE.ER MTN..K..ED S...K..AL. .......... .YI...G.G. I......G.. ........I.  200
Gac1C    .L.Y...G.. .NL...LIK. ..NRH..... .LMD.TEYER MTS.SED.ES G...K..AL. .T........ .Y....G.S. I......G.. .M......I.  200
Tru1C    .R.Y...A.. .NL...A.K. ..KR...... .LMD.AQYEH VTK..A..ES S...K..AL. VT........ .Y....G... I......V.. ........I.  200
Tni1C    .R.Y...G.. .NL...A.K. ..KR...... .LMD.AQYER V.K..A..ES S...K..AL. VT........ .Y....G.S. I......V.. ........I.  200
```

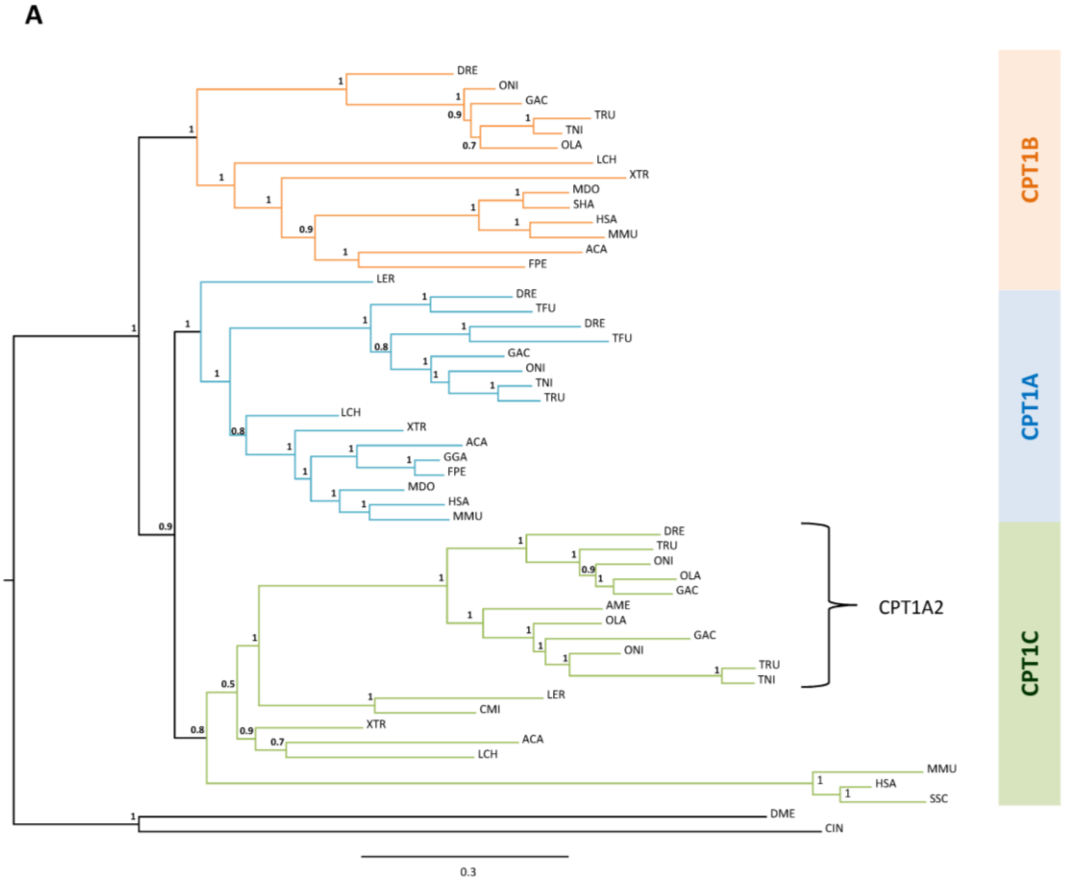
```
Ola1Cb    .L.Y...T.. .HL...A.R. .LTR...... .L.T.PEYKR MTD..N..ES SR..K..AL. .T........ .YI...G.G. I......G.. .......SV.  200
Oni1Ca    .L.Y...S.. .HL...AI.. ..SR...... .L.T.LE.KR MTD..N..ES NR..K..AL. .T........ .Y......G. I......G.. .......SV.  200
Gac_0     .L.Y...T.. .HL...PIK. ..SR..T.A. .L.T.PE.ER MTK..GQ.EA NR.....AL. .T........ .YI...G.G. I......G.. ........V.  200
Tri1C     .L.Y...T.. .HL...AIK. .LSR..R... .L.N.LEYKR MSE..SD.EK NR..K..AL. .T........ .YI...G.G. I......G.. .......SV.  200
Dre1C     .L.Y...T.. .HL...PIK. .LER.....K .L.DLDG.QR MRR.TS..EK SR.....AL. .T........ .Y......S. I......G.. ........N.  200
Dre1B     .L.Y.F.G.. .HL....ID. .I.R...... .L.D..QYKQ METV.ND.KK DKH.K..... .T........ .YI...G.D. I....F.T.. L...I..YR.  200
Oni1B     .L.Y.F.A.. ..L.....D. .IHR...... .L.DN.QYNK MEL..SD.KE NRC.I..... .T........ .YI...G.S. I....F.I.. L......HR.  200
Gac1B     .L.Y.F.A.. ..L.....D. .IHR.....H .L.NSDQY.Q MER..ND.KD SR..I..... GT........ .YI...G.S. I....F.I.. L..I...HR.  200
Ola1B     .L.Y.F.A.. ..L...R.D. .I.R...... .L.DK.QYSQ MET..ND.KE SR..I..... .T........ ..I...G.G. I....F.I.. L..I...HR.  200
Tru1B     .L.Y.F.A.. ..L.....D. .IHR...... .L.VSDEY.Q MVT..K..KD SR..I..... .T........ .YI.....S. I....F.I.. L..I...HR.  200
Tni1B     .L.Y.F.A.. ..L.....D. .IHR...... .L.VSGEYNQ MVA..N..KD SR..I..... .T........ .YI.....S. I....F.I.. L......HR.  200
Hsa1B     ...Y.F.T.. .KL...R.SA .IQR...... .L.D..EYYR MEL..K..QD KK..V..... .......... .YI...G.S. .......V.. LVLIKN.DV.  200
Mmu1B     ...Y.F.T.. .KL.....PA .IHR..D... .L.D..AYYR MET..K..QD KK..V..... .T........ .Y......S. .......A.. .VLIKN.NV.  200
Mdo1B     ...Y.F.T.. .KL...K.AA ..HR...... .L.D..QYYR MEM..KD.QE KK..V..... .T........ .YI...G.S. .......V.. .VFTQH.NI.  200
Sha1B     ...YTF.T.. .KL...K.SA .IKR...... .L.D..EYYR MEM..KD.QE RK..V..... .T........ .YI.....G. I......V.. .VLTPH.EV.  200
Aca1B     ...Y.F.T.. .KL...R.A. .IQR...... .L.DE.R.LD MEA..LD.QQ RK..L..... .T........ .YI...G.S. I......V.. ...T...H..  200
Fpe1B     .L.Y.F.T.. .KL...P.EA .ITR...... .LMD..KYSK MEA..K..KE KK..I..... TT........ QYI..HG.S. .......A.. ......SHI.  200
Xtr1B     ...Y.F.M.. .KL...PLE. .IER..Q... .L.D.DK.SE MKI..E..QK DK..H..... .L........ .YI...G.G. I......A.. Y.....STN.  200
LchB      ...Y.F.TC. .HL.....E. .LHR...... .L.N.LQYKR MEA.TIQ.KH QK..L..... .T........ .Y......T. I......A.. L..II.SSV.  200
Dm        .G.Y.F.G.. ..L.L...K. .MTR..R... .L.D..NYTR MER..K..EQ T...I..... ST........ .Y....G.S. .....F.GT. AIFMNL.DK.  200
ci        NVM.LLVCLI TKTNIIK.EG LGI..YKELK RFIYEKSYEK NVLRSLH.TS NKL.ITNPDF I.IHS.NR.. QY...AG.G. I......G.. L..HN...V.  200

Hsa1C     AARAGNAVHA LLLYRHRLNR QEIPPTLGMR PLCSAQYEKI FNTTRIPGVQ KDYIRDSQHV AVFHRGRFFR MGTHSRNLSP RALEQQFQRI LDDPSPACPH  300
SSc1C     .......... .......... ...F...... .......... .........H R.H.H..R.. .......... V....QS... .......... ..........  300
Mmu1C     .........T .....L.... ...S...... ........RM .........E ..HL...R.. .......... V....P.... ........D. .........L  300
Hsa1A     .......I.. I....RK.D. E..K.IR.TI ......W.RM ...S....EE T.T.Q..K.I V.Y....Y.K VWLYHDG.K. .EM...M... ..NT.EPQ.G  300
Mmu1A     ......TI.. I....RTVD. E.LK.IR.TI ......W.RL ...S....EE T.T.Q..R.I V.Y....Y.K VWLYHDG.R. .E....M.Q. ...T.EPQ.G  300
Mdo1A     .......I.S I....KK.D. E.LK.I..TV ......W.RM ...S....EE T.T.Q..R.I V.Y....Y.K VWLYHDG.K. .EI...M... .....EPQAG  300
Gga1A     ......II.. I....KK.D. ...K.I..TV ......W.RM ...S....EE S.TLQ..K.I V.Y.K..Y.K VWLYHDG.K. .EI...I... ...D.EPQAG  300
Fpe1A     ......VI.. I....KK.D. ...K.I..TV ......W.RM ...S.....E S.VLQ..K.I V.Y.K..Y.K VWLYHDG.K. .EI...I... .N.D.EPQAG  300
Aca1A     .......I.. I....RK.D. .Q.Q.I..TV ......W.RM ...S....EE T.T.Q..K.I V.Y.K.C.YK VWLYYDG.K. .EI...M.W. ...K.KPQ.G  300
Xtr1A     .......I.. I....RK.D. E..K.IQ.TV ......W.RM Y..S....EE T.T.Q..K.I V.Y.K..Y.K VWLYHDG.K. .EI.H.M.K. I..T.SPQ.G  300
Lch1A     ......TI.. I....RK.D. E..K.LMNTI .M..S...RM ...S.....E T.TLQ..K.I V.Y.K..Y.K VWLYHDG.K. .EI...M... ...D.KPQ.G  300
Ler1A     ......TI.. M....RK.D. E..H.LM.QI .M..S...RM ..SS....LE T.TMQ..K.. V.Y.K..Y.K VWLYHSG... CEIQL.M.K. ...Q.LPQTG  300
Xtr1C     ....A.T... I....RKVT. E.LK.LMDCL .M......HM .........D T.T.Q..K.I V.Y.K..... VWVYQGG.N. KE..L...N. .N....PQ.G  300
Ler1C     ......L... .....R..A. EQVQ.ST.PI ......W.RM ..S..L..EE T.RLQ..K.. M.Y.K..Y.K VWLYQSGQ.. SE.QK...Y. ...A.IPQ.G  300
Cmi1C     ......LT.. .....RK.T. E..K.ST.PV ....S.W.RM ..S..T..QE T.HLQ..K.I V.Y.K..Y.K VWVYQGG.R. .E..V..... .A.T.LPQ.G  300
Aca1C     ......L.Y. MMM..RK.VQ E..K.M..CL .M....W.RM ........ME G.S.Q..R.I ..Y.A..... VLLYHNG.R. .E.QA...S. .G..T.PS.G  300
Lch1C     T......IY. C.Q..RKVT. E.LK.L.DCL .M....H.HM ...S....IE T.TLQ..K.I V.Y.K..... VWVYHGG.K. .E..I.IEK. .A.K.SPL.G  300
Tni1A     .......I.. IM...RK.D. AQ.K.LMNTI .M......RM .....V...E T.TLQETK.I V.Y.K....K VWMFYDG.L. .EI...MEK. .A.Q.APQ.G  300
Tru1A     .......I.. IM...RK.D. AQ.K.I.NKV ......W.RM .....V...E T.TLQETK.I V.Y.K....K VWVFYDG.L. .EI...ME.. .A.Q.EPL.G  300
Oni1A     .......I.S IM...RK.D. AQ.K.LMHTI .M......RM .....V...E T.TLQE.K.I V.Y.K..Y.K VWMFYDG.L. .EI...ME.. .A.K.EPL.G  300
Gac1A     .......I.. IM...RK.D. AQ.K.I.NKV ......W.PM .....V..LE T.ILQ..K.I ..Y.K....K VWMFYDG.L. .EI...MA.. .A.TTEPM.G  300
Dre1Aa    ......SI.. MMM..RK.D. AQ.K.LMNTI .M..S...RM ...S.V...E T.VLQE.K.I ..Y.K...YK VWMFYDG.L. .EI...ME.. .A.K.EPQ.G  300
Tfu1A2    V......I.S IM...RK.D. AQ.K.LMNTI .M..S...RM .....V...E E.F.KE.K.I V.Y......K VWMFYDG.L. .EI...ME.. .A.T.MPQ.G  300
Dre1Ab    ......VI.. IM...RK.D. AQ.K.LMNTI .M..S...RM ...S....IE T.SVQ..R.I V.Y....Y.K VWMFYDG.L. .EI...ME.. .A.T.EPQ.G  300
Tfu1A1    ......TI.S IM...RK.D. AQ.K.L.NTI .M..S...RM ...S....IE T.T.Q..R.I V.Y.K..Y.K VSMFYDG.L. .EI...IE.. .A.T.EPQ.G  300
Ola1Ca    ......SI.. FF...RK..K E..K.SR.VI ...A..C.R. ........EE T.TVQ..DY. ..Y.K..Y.. LRVYQAG... .EI.F.I... ....A.PSKG  300
Oni1Cb    ......SI.S FF...RK..K E.LK.W.SAV .C..Y.F.RM .D.C....IL T.TVQ..DYI V.Y.K..Y.. LRVYQAG... .EI.F.I... ......PSKG  300
Ame1C     ......TL.. V.M..R...K E..K.SR.FI ...A..C.R. ........EE T.TVV..EY. ..Y....Y.. LWLYQAG... .EI.Y.I... ......PA.G  300
Gac1C     ......SI.. YF...RK..K E.LK..R.VI ......C.RM ........EE T.TVQ..DYI ..Y....Y.. LRMYHAG... .EI.S.I.K. ......PSKG  300
Tru1C     ......TIY. M....CK..K E..K.PARAV .C..Y.F.RM ...C....TL T.T.Q..DFI V.Y.K..Y.Q LYVYQDG.C. .EI.F.I... .......SKG  300
Tni1C     ......TIY. M....SK..K E..K.PARAV .C..Y.F.RM ...C....TL T.TVH..ECI V.Y.K....Q LCVYQEG.C. .EI.F.I... .....APSKG  300
Ola1Cb    ......TIT. .....RMV.. E.LT.SR.VI ...A..C.RM .....T...E T.VLQ..EF. ..YS...YY. LWVYRAG..A .EI.H.I.W. ......PL.G  300
Oni1Ca    ......TIT. .....RKV.. E.LK.SR.VI ...A..C.RM .....T...E T.VLQ..EF. ..Y....Y.. LWVYRAG... .EI.Y.I... .......L.G  300
Gac_0     ......TIT. .....RKV.. E.LK.SR.VI ...A..C.RM .....T...E T.VLQ..EF. ..Y....... LWVYLAG... .E..H.I... ......PQ.G  300
Tri1C     ......TIT. .....RKV.. E.LK.LWCVI ...A..C.RM .....T..EE T.VLQ..EF. ..Y.K..... LWVYRAGM.. .E..Y.I.K. ......PQ.G  300
Dre1C     ......TIT. .F...RKV.. E.LN.SR.VI ...A..C.RM .....T..EE T.VLQ..EF. ..Y....Y.. LWVYRAG... .EIQF.I... ......PS.G  300
Dre1B     ......V... M.Q..RK.E. G.LT.LR.IV .M..F...RM ........IE T.FVQ.RK.L V.Y....L.K VWLYYGG.W. SE..L..... ...K.EPQ.G  300
Oni1B     ......V... M.Q..RK.E. G.LA.LR.TV .M..T.M.RM ........IE T.FVQ.RK.L V...K....Q VWLYTGG.L. SE..T..... .N.T.EPQ.G  300
Gac1B     ......V... M.Q..RK.E. G.HA.LR.TV .M..T.M.R. ........IE T.IVQ.RK.L V.Y.K....L LWLYTGG.L. SE..T..... .N.T.EPQ.G  300
Ola1B     ......I... M.Q..RK.E. G.HA.LR.TV .M..T.M.RM ........IE T.VVQ.RK.L I.Y.K....Q VWLYTGG.L. SE..M..... .N.TTEPQ.G  300
Tru1B     ......M... M.Q..RK.E. G.HA.LR.TV .M..T.M.RM .....L..IE T.AVL.RK.L I.Y.K....Q VWLYTGG.L. SE..L..... .N.T.EPQQG  300
Tni1B     ......T... M.Q..RK.E. G.HA.LR.TV .M..T.M.RM ........IE T.VVQ.RK.L I.Y.K....Q VWLYTGG.L. SE..L..... .S.T.EPQQG  300
Hsa1B     ...L..II.. MIM..RK.D. E..K.VM.IV .M..Y.M.RM ........KD T.VLQ..R.. ..Y.K....K LWLYEGA.K. QD..M..... ......PQ.G  300
Mmu1B     ...L...... MIM..RK.D. E..K.VM..V .M..Y.M.RM ........KE T.LLQE.R.. ..Y.K....K VWLYEGS.K. .D..M..... ......PQ.G  300
Mdo1B     ...L.SV... MIM..RK.D. E..K.VM.IV .M..Y.M.RM .....L..KD T.VLQ..R.. ..Y.K...YK LWLYQGS.K. .D..M..... ......PQ.G  300
Sha1B     ...L..V... MIM..RK.D. E..K.VM.IV .M..Y.M.RM ........KE S.VLQ..R.. ..Y.K...YK VWLYQGT.K. .D..M..... ...TC.IQ.G  300
Aca1B     .........S I....R..D. EDLA.VM.VV ....Y.M.R. ........K. A.RLL..R.L ....K....K VWLYHAG.P. .D..M..... ......PETG  300
Fpe1B     ......M... I.M..RK.D. G....MM.IV .M..Y.S.RM ........KE T.TLL..K.L ..Y.K...YK VWLYYGG.Q. CD..L..... ......PQ.G  300
Xtr1B     ......VI.. M....RK.E. GL...VM.IV .M..N.MVRM .....V...E T.CLQE.R.. C.Y.K..YY. LALYENG.T. .Q.QA.I.Y. ...S..PQ.G  300
LchB      ......T... M....RK.D. E..K.MMKLV .M..N.V.RM ........LE TGTVKQ.AL. YIYNKAV.VD CMLY.LS.VK LDIL.KQK.. KIIT.YIKYE  300
Dm        ....A.VISL ..NF.RLIEH ..LQ.IM..I ....W...RT ...A.V..LE T.R.I..N.I V.L.K.CYYK .LIYYKG.R. CE.QV.IEE. .KGKATPVEG  300
ci        .S..A.I... MFAF.ST.DK EL.K.IK.VV ........RV ..SC.T...E A.R.C.CR.I ..Y.Q..W.K .TCYKNG.E. SEM.I.IES. .N.T..P.EG  300

Hsa1C     EEHLAALTAA PRGTWAQVRT SLKTQAAEAL EAVEGAAFFV SLDAEPAGLT REDPASLDAY AHALLAGRGH DRWFDKSFTL IVFSNGKLGL SVEHSWADCP  400
SSc1C     .......... ..DM.....K ......E... .......... ...S....DA SG.S...... .......... ........S. .I........ ....A.....  400
Mmu1C     .......... ..SM.....E .V..H..T.. .......... ...S...... .......... .......... .......... .......... ..........  400
Hsa1A     .AR......G D.VP..RC.Q AYFGRGKQS. D...K..... T..ETEE.YR S...T.M.S. .KS..H..CY .........F V..K...M.. NA......A.  400
Mmu1A     .AK....... D.VP..KC.Q TYFARGKQS. D...K..... T..ESEQ.YR E.....I.S. .KS..H..CF .......I.F V..K.S.I.I NA......A.  400
Mdo1A     ..K......G D.VP..KA.Q TYFSRGKQS. D...K..... TM.DTEQ.YS KK..T.M.S. .KS..H.KCY ......T..F ...K...M.. NA......A.  400
Gga1A     ..K......G D.VP..KA.Q AYFSRGKQS. D...K..... T..DDEQ.YS K...S..... .KS.IH..CY ......T... V..K...M.. NA......A.  400
Fpe1A     ..K......G D.VP..KA.Q AYFSRGKQS. D...K..... T..DIEQ.YR KD..K..... .KS.IH..CY ......T... ...K..RI.. NA......A.  400
Aca1A     ..K......G D.VP..KA.Q TYFARGKQS. D.I.K..... T..DTAQ.YR E...TTMET. .KS..H.KCY .......... ...K...M.. NT......A.  400
Xtr1A     ..K......G D.VP..KA.K AYFANGKQSM D...K..... T..ETEQ.YN K...N...S. .KS..H.KCY ......TMSF V..K...M.M N.......A.  400
Lch1A     ..K......G D.VP..KA.. TYFCRGKLS. D...K..... T..DTEQ.FR K...T...R. .KS..H.KCY .......LSF V..K...M.. NS......A.  400
Ler1A     ..K......G D.VP..KA.Q .YFS.GKLS. ..I.K..... T..DTEQ.FR K.E.S...N. .KS..H.KCY .......LSF .I.K...I.. NA......A.  400
Xtr1C     ..K......G E.TA..KA.K TYFRSGLQ.. DL..R..... T.QDDEE..R T...N..... GKS..H.KCY .........F ...K...I.. NA......A.  400
```

```
Ler1C     .KY......G K.IP..K..K .YFSSGKT.M DS..K..... T..EDTPE.F VDNQK...Q. .KS..H.KCY .........F ...A...V.. NA......A.  400
Cmi1C     .........G N.IP.GKA.K .FFSNGRSS. DC..K....L T..GDKP..Q V...K..... .KL..H.KCY .......... ...E...V.. NA......A.  400
Aca1C     ..K.P....G E.DP..RA.N AFFQTGEQS. SI..K..... T..TSEQ..R EPN.Q..... .KS..H..CC .......... ..YR...S.. NA......A.  400
Lch1C     ....P....G D.VP..KA.R DYFQSGRQS. DL..K..... ...ESEQ..K TD..K..... .KL..H.KCY .......... V..K...... NA......A.  400
Tni1A     ..K......G T.TP..NA.D TYFSRGKQ.. D.I.K..... T..DTEQRYD TNN.V...S. .KC..H.KCY ........N. ...K..TM.. NA......A.  400
Tru1A     ..R......G D.TP..NA.D TYFSRGKQS. D.I.K..... T..DTEQCYD TNN.T...S. .KS..H.KCY ........N. ...K..TM.. NA......A.  400
Oni1A     ..R......G D.TP..KA.E .FFSRGKQS. D...K....L T..DTEQRYD TKN.K...I. .KS..H.KCY .......LNM ..YK..TM.. NA......A.  400
Gac1A     ..K......G D.TP..NA.E TYFSRGKQS. D.I.K...C. T..DTEQRFE SDN.Q..VS. .KS..H.KCY ........N. .IYK..TM.. NA......A.  400
Dre1Aa    ..F......G D.VP..KA.S QFFIRGKQS. D...K..... T..DSEQRYE PDN.Q...S. GKS..H.KCY .......LN. ...K..TM.. NA......A.  400
Tfu1A2    ..T......G D.VP..KA.. EFFSTGRKS. D...R..... T..DTEQRYE PDN.Q...S. .KS..H.KCY ........N. .I.K..TM.. NA..T...A.  400
Dre1Ab    ..T......G D.VP..CA.N AYLRHGKKS. DS..K..... T..DTEQRFD QKN.E...R. .KS..H.KCY .......IN. .I.K..TM.. NA......A.  400
Tfu1A1    ..K......G D.VP..CA.D AYLR.GRQS. D...K..... T..DTEQRHN SDS.E..RSF GKS..H.KCY ........N. .IYK.ATI.. NA......A.  400
Ola1Ca    .AK.G..... D.VS..EA.V KYFSSGKRS. DVI.R..... T..D.EQ.TM .D.Q....N. .KS..H.KCY ........SV .YYK...S.I NG......A.  400
Oni1Cb    .AK.G....G D.IP..KA.. KYFSSGKRS. DCI.K..... T..D.EQ.MM GD......R. .KS..H.KCY .........V VYYK...N.I NA......A.  400
Ame1C     ..K.G.F..G D.IP..KA.K EFFSSGKRS. DCI.K..... T..DDEQ.MM GD...NV.R. .KS..H.KCY ........SV VI.K...N.. NA......A.  400
Gac1C     .AK.G....G D.IP..KA.A KYFSSGKRS. DFI.K..... T..DDEQ.GV AD..TRI.S. .KS..H.KCY ........SV VY.K...M.. NG......A.  400
Tru1C     .AR.G....G D.IP..KA.A KHFNSGKKS. DCI.K..... T..D.EQSIV GDNLE...C. IKS..H.KCY N.......SV VFYK...S.. NG....G.A.  400
Tni1C     .AK.G....G D.TP..RA.A KYFSSGKKS. DCI.K..... T..D.EQ.IM GDNLE...H. IKS..H.KCY K.......SV VFYK...S.. NG....G.A.  400
Ola1Cb    ..K.G....G D.VP...I.K EHFSSGKRS. DII.K..... T..D.AQ.MK GD..GN..R. .KS..H.KCY ........SI VIYK...S.. NA......A.  400
Oni1Ca    ..K.G....G D.IP...M.K QYFSSGKRS. D.I.R..... T..D.EQ.MR GD..GN..S. .KS..H.KCY ........SV VIYK...S.. NA......A.  400
Gac_0     ..K.G....G E.VP.F.M.E RHFSSGKRS. DCI.R..... T..D.EQ.MR G...GN..R. .KS..H.KCY ........SI VIYK...S.. NA......A.  400
Tri1C     ..K.G....G D.IP.....K QYFSSGKRS. DVI.K....I T..D.EQ.MR G...GN..R. .KS..H.KCY ........SI VIYK...N.. NA......A.  400
Dre1C     ..K.G....G N.TP..R..K QFFSSGKQS. DCI.K..... T..DQAE.MK G.N.EN..R. .KS..H.KCY ........SV V.YK...N.. NA......A.  400
Dre1B     .LK.PS...G N.VP..RA.L KYFGEGRAS. ..I.T....L T..D.AH.YD P.NIR...L. .KS..H.KCY ........N. ..YK...M.V NT......S.  400
Oni1B     .LK......G Y.IP...A.I KYFS.GKVS. D.I.S....L T..D..Q.YD PAKSN...S. .KS..H.KCY .......... .SYP...M.V N.......A.  400
Gac1B     .LK......G H.VP...S.I KYFG.GKVS. D.I.S....L T..D..Q.YD PAKAK...S. .KS..H.KCY .......... .SYP...V.V NA......A.  400
Ola1B     .LK......G N.VP...A.I KHFSHGKTS. D.I.S....L T..D.SQ.YD GVKKN...S. .KS..H.KCY .......... .SYP...M.V N.......A.  400
Tru1B     .LK......G N.VP..RA.S KYFS.GKVS. D.I.S....L T..D..Q.YD QAR.R...S. .KS..H.KCY .......... .SYP...M.I N.......A.  400
Tni1B     .LK......G N.VP...A.A KYF..GKAS. D.I.S....L T..D..Q.YD HARSR...S. .KS..H.KCY .......... .SYP...M.I N.......A.  400
Hsa1B     ..K......G G.VE...A.Q AFFSSGKA.. ..I.R..... A..E.SYSYD P..E...SL. GK...H.NCY N......... .S.K..Q... NA..A...A.  400
Mmu1B     ..K......G G.VE..EA.Q TFFSSGKMS. D.I.R..... T..EDSHCYN PD.ET..SL. GK...H.NCY N......... .SCK..L... NT......A.  400
Mdo1B     ..K...F..G G.VQ..EA.Q TYFNTGKAS. ..I.K..... T..E.SH..D P.NE...SL. GKS..H.NCY N.......N. .S.K.A.... NT......A.  400
Sha1B     ..K......G G.VQ..EA.Q TYFNTGKAS. ..I.K..... T..E.SH.YD P..E...SL. GK...H.NCY N......... VA.K...... NT..A...A.  400
Aca1B     ..R......G E.LP..EA.E KYFSRGKAS. DC..R..... T..E.EH.FD PDKED...R. SKS..H.QCC ........S. V.YR.....A NA......A.  400
Fpe1B     ..R......G E.VP..EA.A RFFSHGKVS. D.I.R....L T..E.EH.YV AGKEGCM.T. .KS..H.QCY .......... V.YK.....A NA......A.  400
Xtr1B     ..K......G N.VH...A.. NFFSNGRT.. SC..R.V..I ...E.E..YN E..KS..S.. SK...H.NCY N.......S. V..R...... NA......A.  400
LchB      RKGSVRISLS C.IP..KA.S EFFSHGKIS. T.I.R....M T..D.EQAYD K.N.TT..S. .KS..H.KCF .........F ...K.....I NT......A.  400
Dm        .........W N.SK..EA.N TFFSWGQTS. RTI.S...VL ...D..FEFD LAR.EL..NF GKK..H.N.Y N.....C..V C.GT..RV.F NA..T.S.AA  400
ci        .........G E.IP..KA.N TYFVDGKKS. H.I.K...IL V..D.EHVVS D..S...SK. GRS..H.KCY N.....T.NC ...K..RW.I NA......A.  400

Hsa1C     ISGHMWEFTL ATECQLGYST DGHCKGHPDP TLPQPQRLQW DLPDIHSSIS LALRGAKILS ENVDCHVVPF SLFGKSFIRR CHLSSDSFIQ IALQLAHFRD  500
SSc1C     .......... .........A .......... S.......H. .....L.... ......QA.A ..I....F.. .H......K. .......... T.........  500
Mmu1C     V...L..... ........A. .......... .......... ...E.QP... .......T.. G.I....F.. .H......KC ..V....... LV........  500
Hsa1A     .VA.L..YVM SIDS....AE ......DIN. NI.Y.T.... .I.GCQEV.E TS.NT.NL.A ND..F.SF.. VA...GI.KK .RT.P.A.V. L......YK.  500
Mmu1A     .V..L..YVM ..DV.....E ......DKN. NI.K.T.... .I.GCQEV.E TS.SS.SF.A ND..L.SF.. DT...GL.KK .RT.P.A... L......YK.  500
Mdo1A     .V..L..YVM ..D.....TE ......DTN. NI.Y.T.... EI.ECQDV.E ES.SL.ST.A ND..F.SF.. DA...EL.KK SRT.P.A.V. L......YK.  500
Gga1A     .V..L..NVM ...YE...LE ......DTNQ NI.I.TK... EI.ECQDV.E RS.ST.RA.A DD..FYSFY. DV...GL.KK AKT.P.A... L......Y..  500
Fpe1A     .V..L..NVM ...YE....E ......DINQ NI.I.TK... EI.ACQEV.E RS.ST.IA.A DD..FYSFF. DA...GL.KK AKT.P.A.V. L......Y..  500
Aca1A     .V..L..NVM FSD.E...TE ......ESSS GILM.S.... EILECQEV.E RS.AV.RP.A DD..F.SF.. DT...GLMKK AKT.P.A.V. L......Y..  500
Xtr1A     .V..L..YVM ..DKE...NE ......DVNG NI.P.S.... .I.ECQNVVE ES.TV..A.A DD..F.SF.. NS...GL.KK SRT.P.A.V. LS.....Y..  500
Lch1A     .V..L..YV. ..DS....TE E.....E.K. SI.F....R. EI.ECQEV.E IS.KV..A.A DD..F.SF.. DS...GL.KK .RT.P.A... L......YW.  500
Ler1A     .I..L..YV. ..DQ....TD E.N...E.N. QIQP...... .IEECQEV.H QS.SV.QQ.A DD..F.SF.. DK...GV.KK .WI.P.A.V. M......Y..  500
Xtr1C     .V......V. ..D.....NE E.N...QV.S N..V...... EISECQEV.Q SS.AV.QA.A DD..F.TF.. KE...GL.KK .RT.P.A... .......Y..  500
Ler1C     .I..L..YA. ..DT....KE ..N...D.A. NV.L...... .I.KCQEV.M SS.KV.QT.A ND..FYAF.. TR...GV.KK .RT.P.A... ......NY..  500
Cmi1C     .I..L..YV. ..DS....ND Q.....DAES .VLS...... .I.ECQEV.. GS.KV.QS.A ND..F.TF.. TK...GL.KK .RT.P.A... L.....N...  500
Aca1C     .V..L..YC. ..DAT...DA Y.N...DM.. NV.P..K... EI.PCEAV.M QSF.V.YN.A SDI.F.AFT. KD...GL.KK .RT.P.G... L.........  500
Lch1C     .I..L..YV. ..DT.I..KP ......EA.S .ILP...... .I.ACREV.K MS.AV.QT.A ND..F..FT. KD...GL.KK .RT.P.A... .......Y..  500
Tni1A     .V..L..QV. SMDPN...TE ....R.A.H. N..G...... .ISTCQQV.Q SS.TV.QK.A DD..S.II.. .E...GK.KK .RT.P.A... .......Y..  500
Tru1A     .V..L..HV. SMDPN...TE E...R.V.H. N..G..K... .I.ACQQV.Q NS.TV.QN.A DD..S.II.. .E...GK.KK .RT.P.A... .......Y..  500
Oni1A     .V..L..HV. SMDPK...TE ....V.K.H. N..G...... TI.ACQEA.E SS.TV.RA.A DD....II.. TD...GL.KK .RT.P.A... .......Y..  500
Gac1A     .V..L..HV. SMDPK...TE A.....E.H. N..G..K.S. .I.ACQEV.Q SS.KV.RT.A DD..S.IM.. .D...GL.KK .RT.P.A... .......Y..  500
Dre1Aa    .V..L..QV. SSDPR...TE E.....N.H. NM.G...... .I.ECQTV.. SS.KV.NT.A DD..M.IF.. ND...GL.KK .KT.P.G... L.........  500
Tfu1A2    .V..L..HV. SMDPT...TE ....R.K.H. N..G.L.... .ISVCQ.V.R SS.KV.NA.A DD..M.IF.. NN...GL.KK .KT.P.G... ..........  500
Dre1Ab    .V..L..QV. SMDPK...TE ......E.HA N..G...... NI.TCQTM.T NS.SV.EA.A DD..S.II.. .D...GL.KK .RT.P.A... L......Y..  500
Tfu1A1    .I..L..NV. S.DAK...TD ....A.QTHR N..G...... NI.PCQTM.A NS.SV.QA.A DD..MVII.. QE...GL.KK .KI.P.A... ......N.K.  500
Ola1Ca    VVA.V..YV. ..DS....NE E.....EV.A S.....K.N. EISPCEEQ.. RS.AV.QA.A DD..F..LS. RD...GQ.KK .KV.P.A... MT....YY.E  500
Oni1Cb    VLA.V..Y.. .NDS....NA E.....DV.. S..R.VK.S. EI.PCEEQ.A QS.AV.QA.A DD..F..FS. QE...GKVKK .RV.P.G... M.....Y..E  500
Ame1C     .VS....YA. ..DS....NE E.....DVN. S.......T. .I.KCQEQVA QC.AV.QP.A DDI.F..FA. RD...GS.KK IKM.P.G... L.....FY..  500
Gac1C     VLS.A.QYV. T.D.....NA E.....EV.S S..G..K.N. EI.PCEEQ.. GS.AV.QA.A DD..V..FS. EE...GK.KK .RV.P.A... LT.....Y.N  500
Tru1C     VLT.L..Y.. ........NA E.....EV.A S.AE..K.N. EISSCEEQ.. QS.EV.QA.A ND..M...V. RD...AKVKK .RVNP.A... ......YY..  500
Tni1C     VLS.L..Y.. ........NA E.....EV.A S..K..K.N. EI.SCEEQ.C QS.AL.QA.A ND..M...V. RD...AKVKK .RVNP.A... .......Y..  500
Ola1Cb    TVA.L..Y.. ..DA....TE ......EV.R S..P.H..S. EI.SVQDQ.F SS.TL..A.A DD.....F.. RD...GK.KK LRV.P.A... .S....Y...  500
Oni1Ca    TVA.L..Y.. ..DA....TE ......EVE. S.......V. NI.AVQAQV. SS.AV.QA.A DD.....F.. RE...GR.KK .RI.P.A... .S....YY..  500
Gac_0     TVA.L..Y.. ..DA....TE ......DV.R S..P....A. .I.SVQAQA. SS.VV.QA.A DD.....F.. RD...GR.KK LRV.P.A... .G....Y...  500
Tri1C     TVA.L..Y.. ..DAH...TE ......EVE. L..H....L. .I.LCNTQVQ SS.AV.QA.A DD....I.H. RD...GR.KK LRV.P.A.V. .S....YY..  500
Dre1C     .VA.L..... ..DTH...NS ..N.R.DV.H S..H....S. .I.FVQTQ.. ES.AV.QA.A DE.....F.. RK...GL.KK MK..P...V. L.....YY..  500
Dre1B     .I.....YV. ..D.H...TA E.....DVNK ..AP.T.... .I.KCQEI.E GSY.I..GIA DD..F.GCL. NE...GL.KK .RT.P.A... L.....QY..  500
Oni1B     .V.....YI. ..D.H...TE E.....DVNK N..H.T.... QI.NCQNV.E TSYLS..QIA DD..F.GYL. AE...GL.KK .RT.P.A... L.....Q...  500
Gac1B     .V.....YV. ..D.H...TE E.....DVNK G..H.S.... QI.NCQEV.E TSYLS..LIA DD..F.GYL. HE...GL.KK .RC.P.A... L.....Q...  500
Ola1B     .V.....YV. ..D.H.C..E E.....DANR G..F.T.... QISKCQDV.E ASYLS..KIA DD..F.GHL. TE...GL.KK .RT.P.A... L.....Q...  500
Tru1B     .V.....YV. S.D.H...TE E.....DVNK G..Y.S.... QI.VCK.I.E ASYVS..RIA DD..FYGCL. HE...GL.KK .RT.P.A... M.....Q...  500
Tni1B     VV......V. ..D.H...TE E.....DVNK G..Y.S.... QI.VCQ.I.E ASYVS..QIA DD..F.GCL. HE...GL.KK .RT.P.A... M.....Q...  500
Hsa1B     .I..L...V. G.DSH...TE T...L.K.N. A.AP.T.... .I.KCQAV.E SSYQV..A.A DD.ELYCFQ. LP...GL.KK .RT.P.A.V. ..........  500
Mmu1B     .I..L...V. G.DTH...TE T...V.E.NT ...P....P. .I.ECREA.E NSYQV..A.A DD.ELYCFQ. LP...GL.KK .RT.P.A.V. ..........  500
Mdo1B     .V..L...V. ..DAH.D.TD A...Q.K.NH S.AP....L. .I.ECQEL.E SSYQV..T.A DD.ELYCFH. AT...GL.KK .RT.P.A.V. ..........  500
Sha1B     VV..L...V. ..DAH.D.NE S...Q.K.NH S.AP...... EI.ECQKI.E SSYEV..A.A DD.ELYCFQ. LP...GL.KK .RS.P.A.V. ..........  500
Aca1B     .I..L...M. ..DH....CS ....H.V.NT A..P....T. .I.ECCNV.D ASYAV.RA.A DDI.F.SFR. VE...GR.KK .RT.P..... ..........  500
Fpe1B     .I..L...A. ...K....TD R...R.E.NT Q.AP...... .I.QCRDT.E SSY.L..A.A DD..FCCFQ. .E...GL.KK .RT.P.A... .S........  500
```

```
Xtr1B         .I..L..... ..D.E...TE ..N.R.DAGS P..P.Y.... .I.PCREV.E RSYVT..AIA DD..F.CLC. .D...GL.KK .RS.P.A.F. .......Y.E  500
LchB          VI..L...V. ..D.E.S.TE S.....EMNK K..P...... ...ECKEM.Q QSYKV..A.A DD.NFCCF.. MS...GM.K. FKT.P.A... ..........  500
Dm            .AS....NLI VDDLSD..DE T.NT..T.AF QP.T.T..T. ..KPCLAQ.E E.TIDVTK.I NE.NLRILVH QDY..G.MKK .RI.P.AY.. M.....YY..  500
ci            .MSYVV.EA. GF.YS...TQ ..RV..R.TV QPIT.H.... Q.TPCQEV.E TS.SV.NN.A DD.HLN.SA. KH...GLVKK FKM.P.A... A...I..L..  500

Hsa1C         RGQFCLTYES AMTRLFLEGR TETVRSCTRE ACNFVRAMED KKTDPQCLAL FRVAVDKHQA LLKAAMSGQG VDRHLFALYI VSRFLHLQSP FLTQVHSEQW  600
SSc1C         ..R....... .......... .......... ..S......H Q......... ..L....... .......... .......... ..Q..R.... ..D.......  600
Mmu1C         .......... .......... .......... ..Q.....DN .E..QH.... .......... .......... I......... M..L..M... .....Q.Q..  600
Hsa1A         M.K......A S.....R... ........T. S.D.....V. PQ.VE.R.K. .KL.SE...H MYRL..T.S. I.....C..V ..KY.AVE.. ..KE.L..P.  600
Mmu1A         M.K......A S.....R... ........T. S....L..M. PT.AE.RFK. .KI.CE...H .YRL..T.A. I.....C..V ..KY.AVD.. ..KE.L..P.  600
Mdo1A         M.K......A S.....R... ........M. S....L..VN PESVENK.K. L.I.AE...H MYRL..T.A. I.....C..V ..KY.AVD.. ..KE.L.DP.  600
Gga1A         M.K.S....A S.....R... ........I. S....QT..N PESNENKMKS ..L.AT...H .YRL..T.A. I.....C..V ..KY.SVD.. ..KE.L..P.  600
Fpe1A         M.K.S....A S.....R... ........V. S.....T... PESSENK.KF ..I.AA...H .YRL..T.A. I.....C..V ..KY.AVD.. ..KE.L..P.  600
Aca1A         M.K......A S.....R... ........TQ S.K..K..D. PE.LEKKI.. .KA.ATH..L .YRL..T.N. I.....C..V ..KY.AVE.. ..KE.L..P.  600
Xtr1A         KEK......A S.....R... ........I. S.D..L..S. PQ.NEKR.Q. .KE.AE...Q MYRL..T.S. I.....C..V ..KY.GVD.. ..KE.L..P.  600
Lch1A         K.K.Y....A S.....R... ........V. S....LS.V. PQ.NE.R.K. .KI.SE...N MYRL..T.A. I.....C..V ..KY.GVD.. ..KE.L..P.  600
Ler1A         K.K......A S.....R... ........SQ STQ..L..VN SQ.NEEK.K. ..L.A....S MYRH..T.E. I.....C..V ..KY.GMD.. ..KE.L..P.  600
Xtr1C         K.K......A S.....R... ........V. SSQ..K..MN PQ.NEER.K. ..L.AE...M .YRH..T.G. I.....C..V ..KY.GVD.. ..KE.L..P.  600
Ler1C         K.T....... S.....R... ........V. T....T.LDG .HSNEE..N. ..I.AKN..H ISRL....C. I.....C..V ..KY.GVS.. ..QE.L..P.  600
Cmi1C         K.K....... S.....R... ........IQ SS........ HLSASE..K. .HL.AKT..H ISRLS.A.C. I.....C..V ..KY.GVH.. ..QE.L..P.  600
Aca1C         KKK......A S.....R... ........I. S....K..M. PQD.SAR.H. ....AE...N .YRQ..T.A. I.....C..V ..KY.G.D.. ..RE.L..P.  600
Lch1C         K.K......A S.....R... ........S. S.A..K..V. AY.NAAR.K. .K..AE...Q MYRL..T.A. I.....C..V ..KY.GVD.. ..NE.L..P.  600
Tni1A         K.K......A S.....R... ........V. S.A...S.IR DE.TEER.R. LKK.AE...N .YRL..T... I.....C..V ..KY.GED.A ..KE.L..P.  600
Tru1A         K.K......A S.....R... ........M. S.A...S.IR DE.TEER.R. LKK.AE...N MYRL..T.E. I.....C..V ..KY.GEE.A ..KE.L..P.  600
Oni1A         K.K......A S...M.R... ........I. T.A.....VG DE.REER.R. LKL.AE...N .YRL..T... I.....C..V ..KY.GED.. ..KE.L..P.  600
Gac1A         KKK......A S.....R... ........S. S.A.....IK GE.RDE..R. LKQ.AE...N MYRL..I.E. I.....C..V ..KY.GED.. ..KE.L..P.  600
Dre1Aa        K.K......A S.....R... ........N. T.A..H..M. EA.REER.K. LKA.TE...N .Y.L..T.K. I.....C..L ..KY.GED.. ..KE.L..P.  600
Tfu1A2        KKK......A S.....R... ........M. T.D.....M. EQ.REEK.R. LKL.AE...E .YRM..T.K. I...I.C..V ..KY.GDD.A ..KE.L..P.  600
Dre1Ab        K.K......A S.....R... ........T. S.A.....NS NH.RE.K.Q. LKN.AE...Q MYRL..T.H. I.....C..V .LKY.GQD.. ..KE.L..P.  600
Tfu1A1        K.K......A S.....R... ........I. TSA..K..VN NE.REKK.S. L.H.AGN..H MYQM..T.K. I.....C..V ..QY.QQD.. ..KK.L..P.  600
Ola1Ca        ..T......A S.....R... ........N. SSA....L.N GQPADV.RR. ..G.SE...Q .YRM..T.A. I...I.C..V ..KY.G.E.. ..KE.LA.P.  600
Oni1Cb        ..T......A S.....R... ........N. SSA.I..L.G GEATDV.KC. ....SE...L .YRL..T.A. I.....C..V ..KY.GVE.. ..KE.L..P.  600
Ame1C         ..M......A S.....R... .......SS. S.A.IK.F.N GEDVEV.RR. I.N.SE...M .YRL..T.A. I.....C..V ..KY.GIE.. ..KEAL..P.  600
Gac1C         Q.R......A S.....K... ........N. S.A....L.G GEDADV.RR. .HE.SE...Q .YRM..T.A. I.....C..V ..KY.QVE.. ..KE.L..P.  600
Tru1C         QKR.S..... S.....R... ........N. S.A.I..L.G GE.TDV.RR. ..T.SE...R .CRM.ST.A. I.....C..V ..KC.GVE.. ..KELE..A.  600
Tni1C         QKR.S..... ......R... ........N. S.A.I..L.G GE.TDV.RR. ..T.CE...R .CRM.ST.A. F.....C..V ..KY.GVE.. ..KE.L..A.  600
Ola1Cb        ..S......A S.....R... .......SNQ S.A..K.L.G GEGAD..KR. ..Q.SE...N .YRM..T.A. I.....C..V ..KY.GVD.. ..KE.L..P.  600
Oni1Ca        .SG......A S.....R... .......SN. S.A.IK.L.S GE.EEE.RR. .QL.SE...N .YRM..T.A. I.....C..V ..KY.GVE.. ..KE.L..P.  600
Gac_0         ..G......A S.....R... .......SN. SSA....L.T .EAVD..RH. ..L.SER..N .YRM..T.A. I.....C..V ..KY.GVD.. ..KE.L..P.  600
Tri1C         ..S......A S.....R... ........N. SAA..K.L.N GEDEES.RR. ..L.SET..N .YRM..T.A. I.....C..V ..KY.GVD.. ..KE.L..P.  600
Dre1C         ..T......A S.....R... .......SN. S.A..L.L.G GEDRE..RK. L.KTAE...N .YRL..T.S. I.....C..V ..KY.GVE.. ..KE.L..P.  600
Dre1B         K.E....... S...M.R... ........C. STA....... DT.NE.R... .KQ.AE...N MYRL..T.A. I.....C... ..KVMGID.. ..K..L..P.  600
Oni1B         QRV....... S...M.RD.. ........S. .VA....... AA.NA.R... ..K.AE...N MYRL..T.S. I.....C... ..KY.GVD.. ...K.L..P.  600
Gac1B         Q.V....... S...M.RD.. ........T. .VA....... EA.KA.R... ..K.AE...N MYRL..T.S. I.....C..L ..KY.GVD.. ..KK.L..P.  600
Ola1B         Q.V....... S...M.RD.. ........S. .VA....... VA.NA.R.S. .QK.AE...N MYRL..T.S. I.....C... ..KY.GAD.. ..NK.L..P.  600
Tru1B         Q.L....... S...M.RD.. ........S. .VA..K...S AS.NA.R... ..K.A....N MYRL..T.S. I....LC... I.KY.NVD.. ..KK.L..P.  600
Tni1B         Q.L....... S...M.RD.. ........S. .VE......S AA.NA.R... ..K.A....N MYRL..T.S. I....LC... ..KY.GVD.L ..KK.L..P.  600
Hsa1B         ..K......A S...M.R... ........S. STA..Q..ME GH.KADLRD. .QK.AK...N MYRL..T.A. I.....C..L ..KY.GVS.. ..AE.L..P.  600
Mmu1B         K.K......A S...M.R... ........N. SAA..Q..MK GHKKQDLQD. ..K.SE...N MYRL..T.A. I.....C... ..KY.GVS.. ..AE.L..P.  600
Mdo1B         K.K......A S...M.RD.. ........T. .TA..Q..M. SYMK.DLQD. ..K.AE...L .YRL..T.A. I.....C..V ..KY.G.H.. ..A..L..P.  600
Sha1B         K.N......A S...M.RD.. ........A. TTA.....T. SYMK.DLQD. ..K.AE...H .YRL..T.A. I.....C..V ..KY.G.H.. ..A..L..P.  600
Aca1B         K.R......A S.....R... ........S. .TA...S.A. PCSSSER.E. .H..AE...H .YRL..T.S. L.....C..V ...Y.GVE.. ..DK.L..P.  600
Fpe1B         K.C......A S.....R... ........A. STA...S.G. A..VTERQR. .KL.A....H MYRL..T.A. I.....C..V ...Y.GI... ..A..L..P.  600
Xtr1B         K.H......A S.....RD.. ........TQ TSD..K.... PQSQEKR... Y.A.AEH..L MYRW..T.K. I.....C... ..KY.GTD.A ..QK.L..P.  600
LchB          K.A....... S.....RD.. ........N. T.A..Q..L. T...KER.S. ..K.SE...Y .YRL..T.A. I.....C..V ..K..GVH.. ..N..LA.P.  600
Dm            A.R.S....A S.....R... .....P..I. SSAW.K..QN PT.NDERVKM MQA.C.R..L GYQD..C.R. I.....C..V ..KY.EVD.. ..NE.L..P.  600
ci            K.R.S....A S.....R... ........S. M.A.AKS... DF.NKDRY.. LKK...R.IN GY.E..T... I...I.C..V ..KY.K.E.. ..QK.LQ.P.  600

Hsa1C         QLSTSQIPVQ MDPDYVSSGG GFGPADDHGY GVSYIFMGDG MITFHISSKK SSTKT 655
SSc1C         ..A....... .......... ......E... ......T.ED T......... ...R. 655
Mmu1C         L.....V... T......... .....H.... .I......EN A......... ...E. 655
Hsa1A         R.....T.Q. V..E...... ....VA.D.. .....LV.EN L.N......F .CPE. 655
Mmu1A         R.....T.Q. V......C.. ....VA.D.. .....IV.EN F.H......F ..PE. 655
Mdo1A         R.....T.Q. VN.E...... ....VA.D.. .....IV.EN L.N..V...F ..PE. 655
Gga1A         R.....T.QH I..EML.... ....VA.D.. .....ILDEN S.H..V...F .CSE. 655
Fpe1A         R.....T.QH I..EML.C.. ....VA.D.. .....ILDEN S.H..V...I .CSE. 655
Aca1A         R.....T.QH I..GME.... ....VA.D.. .....IV..N L.N..V...Y .CPE. 655
Xtr1A         R.....T.Q. VQ.EN..... ....VA.D.. .....IV.EN L.N......F ..PE. 655
Lch1A         R.....T.Q. V..E...... ....VA.D.. .....LV.EN L.NL.V...L ..LE. 655
Ler1A         K.....T... ...K...... ....VA.D.. .....MV.EN L.NM.....F ..PE. 655
Xtr1C         R.....T.I. V..E...C.. ....VA.D.. .....IV.EN L.N......F ..HE. 655
Ler1C         C.....T.I. I..E.I.I.. ....VA.D.. .....IV..H L.N....C.V ..PL. 655
Cmi1C         S.....T.L. V..E.I.C.. ....VA.D.. ....FIV.EN L.N....C.V ..QY. 655
Aca1C         R.....T.I. L......C.. ....V..N.. .....IV.ED L.N..V.C.V ..PE. 655
Lch1C         R.....T.I. A..E...C.. ....VA.D.. .....IV.ED L.N....G.I .GLD. 655
Tni1A         R.....T.L. L..E..T... ....VA.D.. .....IL.EN L.N......R ..PE. 655
Tru1A         R.....T... V..E...... ....VA.D.. .....IV.EN L.N......R ..PE. 655
Oni1A         K.....T.L. V..E...... ....VA.D.. .....IL.EN H.N......R ..PE. 655
Gac1A         R.....T.L. V..E...A.. ....VA.D.. .....IL.EN L.N......H ..PE. 655
Dre1Aa        R.....T.L. ...E..T... ....VA.D.. .....IL.ED L.N......H ..HE. 655
Tfu1A2        R.....T.L. I..E..T... ....VA.D.. .....IL.ED L.N......Y ..IE. 655
Dre1Ab        R.....T.L. G..E..T... ....VA.D.. ..A.VIV.EK L.N......R ..PE. 655
Tfu1A1        R.....T.L. P..E...... ....VA.D.. ..A..IV.EN L.N....C.Y ..PEA 655
Ola1Ca        R.....TSI. V..E.I.C.. ....VA.D.. ....N.I.EN V.N....C.H .CPS. 655
Oni1Cb        R.....T.I. V..E...C.. ....VA.D.. ....CVL.EN ..N....C.H .CPD. 655
Ame1C         R.....T.F. L....ITC.. ....VA.D.. ....SLI.EK I.S..V...H .CPD. 655
Gac1C         R.....T.L. V..E...C.. ....VA.D.. ....CAL.EK .LS...TC.H .CPN. 655
Tru1C         R..S.H..YM I..E.LCY.. ....VT.D.. ..C.LML.GD VL.L...C.N .CPD. 655
```

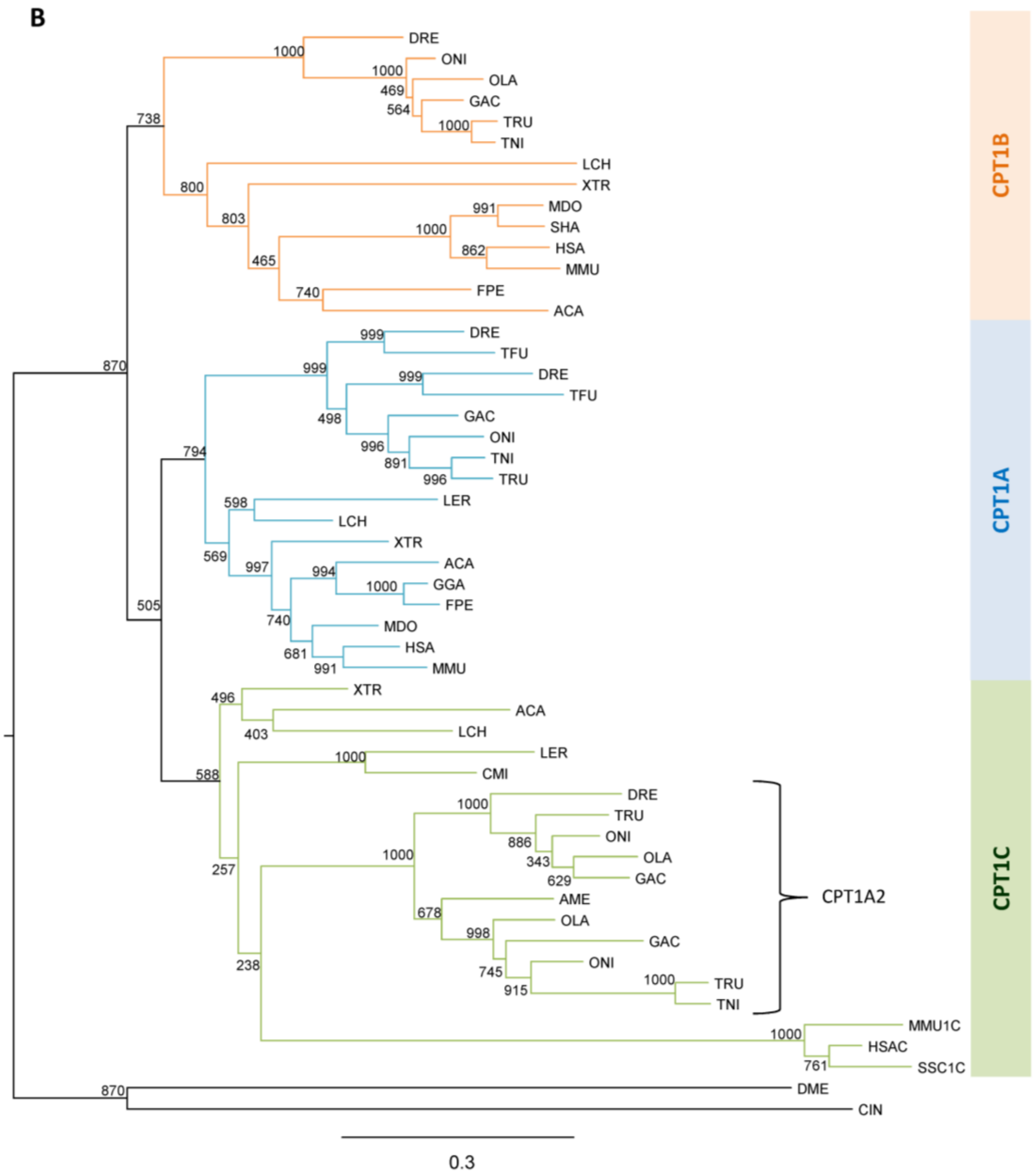
```
Tni1C       ..AS.HV.HM I..E.LCY.. ....VT.G.. ..C.LML.GD VL.L.V.C.N .CPA. 655
Ola1Cb      R.....T... ....F..L.. ....VA.D.. .....IV.ED ..N..V...H .CGE. 655
Oni1Ca      R.....T... ....FI.L.. ....VA.D.. .....IV.ED .VN..V...Y .CSQ. 655
Gac_0       R.....T... ....FI.L.. ....VA.D.. .....IV.ED ..N..V...H .CSE. 655
Tri1C       R.....T.I. I......L.. ....VA.D.. .....IT.ED ..N..V...H .CNQ. 655
Dre1C       R.....T... ...EFI.L.. ....VA.D.. .....I..ED ..N..V.C.H .CKE. 655
Dre1B       R.....T.Q. L..K...A.. ....VA.D.. .....IV.EN L........F ..PE. 655
Oni1B       K.....T.Q. L..K..GG.. ....VA.D.. .....IV.EN L........F ..PD. 655
Gac1B       R.....T.Q. L..K...A.. ....VA.D.. .....IV.EN L........F ..PD. 655
Ola1B       R.....T.Q. L..K..GA.. ....VA.D.. .....IV.EN L........F ..PN. 655
Tru1B       R.....T.Q. L..K..GA.. ....VA.D.. .....IV.EN L........F .CPH. 655
Tni1B       R.....T.Q. L..K..GA.. ....VA.D.. .....II.EN L........F .CPH. 655
Hsa1B       R.......Q. I..NHLGA.. ....VA.D.. ....MIA.EN T.F......F ..SE. 655
Mmu1B       S.......Q. I..NHLGA.. ....VA.D.. ....MIA.EN TMF......Y ..SE. 655
Mdo1B       R......AQ. I..NHIAA.. ....VA.D.. ....MIA.EN T.F..V...F ..SE. 655
Sha1B       R.....TAQ. I..NHLAA.. ....VA.D.. ....MIA.EN T.F......F ..SE. 655
Aca1B       R.....T.Q. I..ECA.... ....VA.D.. .....IA.EN LV...V...F ..PE. 655
Fpe1B       R.....T.Q. L...H..... ....VA.D.. .....IA.EN L....V...F ..SE. 655
Xtr1B       R.....T.Q. L...H..A.. ....VA.D.. ......A.EN L..L.....F ..PE. 655
LchB        A.....T.Q. TN........ ....VAED.. .....IV.ED L........Y ..PE. 655
Dm          R.....T.H. T..NCI.A.. ....VA.D.. .....IA.EN L.F....A.T TCQQ. 655
ci          R.....T.H. A..NFL.G.. ....VA.D.. .....ICHEN L.M..V...Y ..SE. 655
```

**Supporting information 3 Figs:** Supporting Phylogenetic analysis



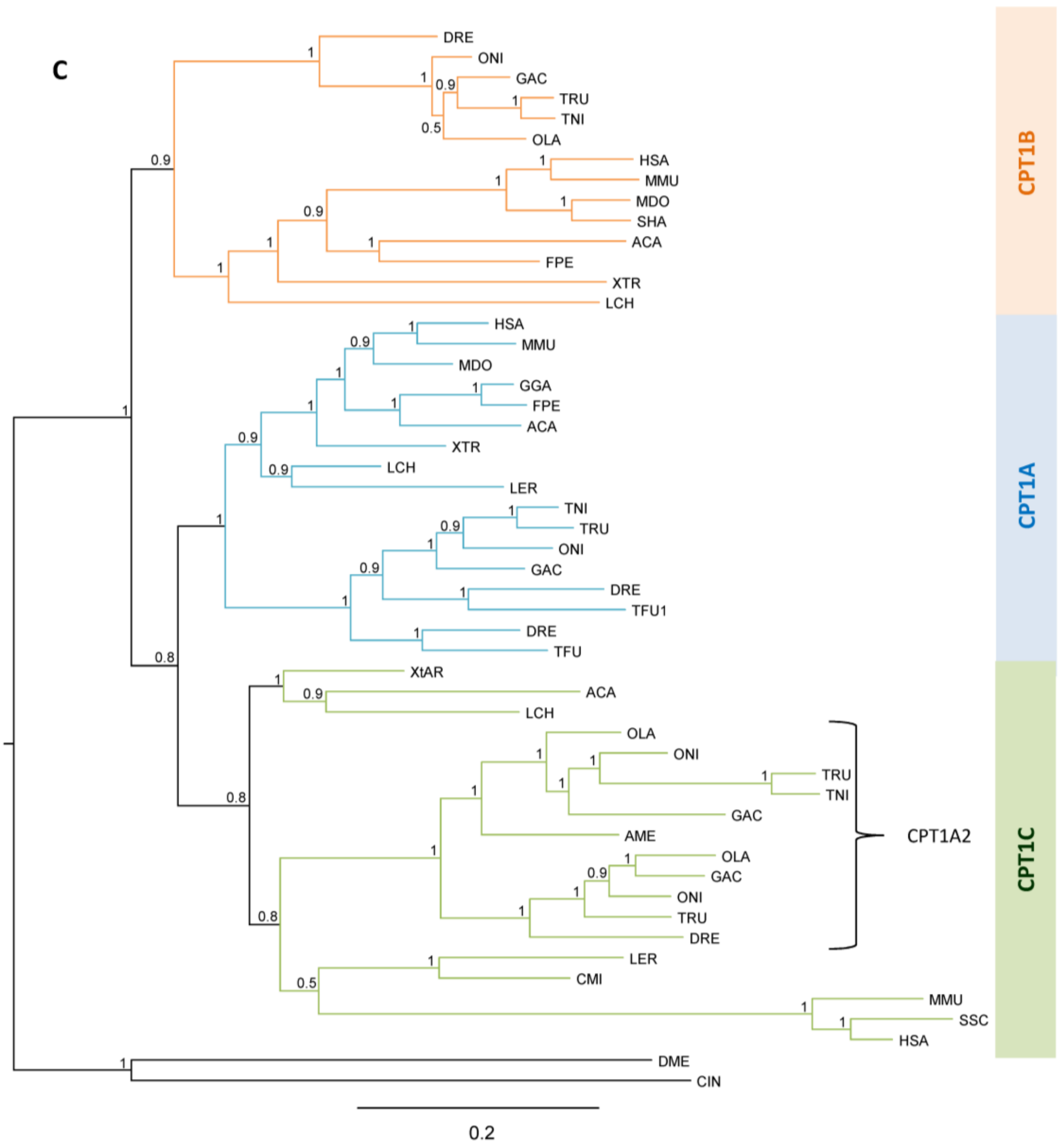
Phylogenetic analysis methods are the same as described in the main manuscript with the exception that sequence alignment input in PhyML contained gaps.

HSA – *H. sapiens*; SSC- *S. scrofa;* MMU – *M. musculus*; SHA – *S. harrisii*; MDO – *M. domestica*; ACA – *A. carolinensis* ; GGA – *G. gallus*; FPE – *F. peregrinus;* XTR – *X. tropicalis* ; LCH – *L. chalumnae* ; DRE – *D. rerio*; AME – *A. mexicanus;* TNI – *T. nigroviridis*. TFU – *T. fulvidraco;* GAC – *G. aculeatus* ; TRU – *T. rubripes* ; ONI – *O. niloticus* ; OLA – *O. latipes* ; CMI – *C. milii;* LER - *L. erinacea;* CIN – *C. instestinalis* ; DME – *D. melanogaster*



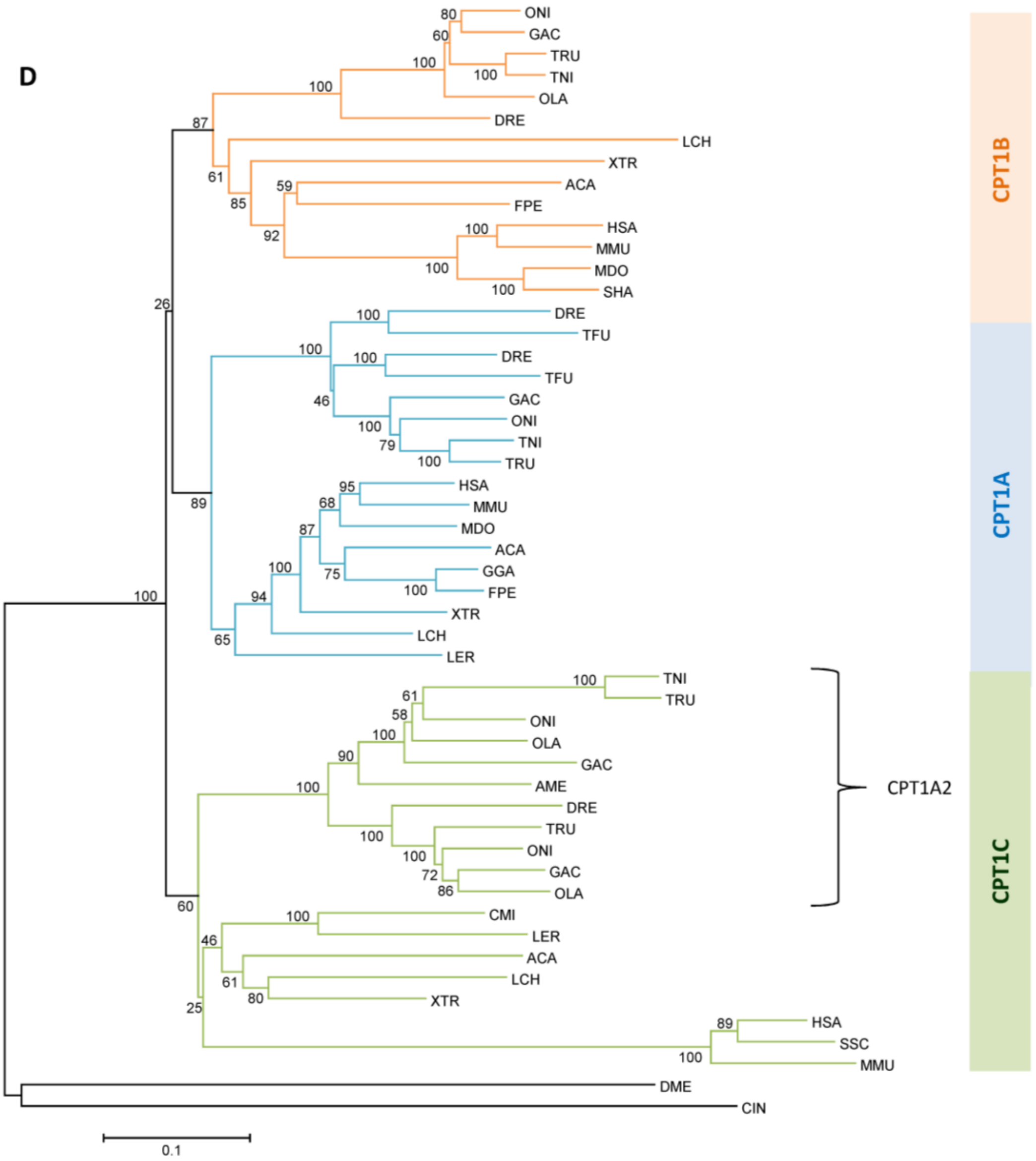
Maximum likelihood phylogenetic analysis was performed in PhyML v3.0 server. using the protein evolutionary model LG +G +F (previously calculated in Protest) and the number of bootstrap replicates was set to 1000. The resulting tree was visualized in Fig Tree V1.3.1 avaliable at http://tree.bio.ed.ac.uk/software/figtree/ and rooted DME and CIN

HSA – *H. sapiens*; SSC- *S. scrofa;* MMU – *M. musculus*; SHA – *S. harrisii*; MDO – *M. domestica*; ACA – *A. carolinensis* ; GGA – *G. gallus*; FPE – *F. peregrinus;* XTR – *X. tropicalis* ; LCH – *L. chalumnae* ; DRE – *D. rerio*; AME – *A. mexicanus; TNI – T. nigroviridis.* TFU – *T. fulvidraco;* GAC – *G. aculeatus* ; TRU – *T. rubripes* ; ONI – *O. niloticus* ; OLA – *O. latipes* ; CMI – *C. milii;* LER - *L. erinacea;* CIN – *C. instestinalis* ; DME – *D. melanogaster*



Bayesian phylogenetic analysis was performed using MrBayes v3.2.3 available in CIPRES Science Gateway V3.3. MrBayes was run for 1 million generations with the following parameters: rate matrix for aa=mixed, nruns=2, nchains=4, temp=0.2, sampling set to 1000 and burin to 0.25. The resulting tree was visualized in Fig Tree V1.3.1 avaliable at http://tree.bio.ed.ac.uk/software/figtree/ and rooted DME and CIN.

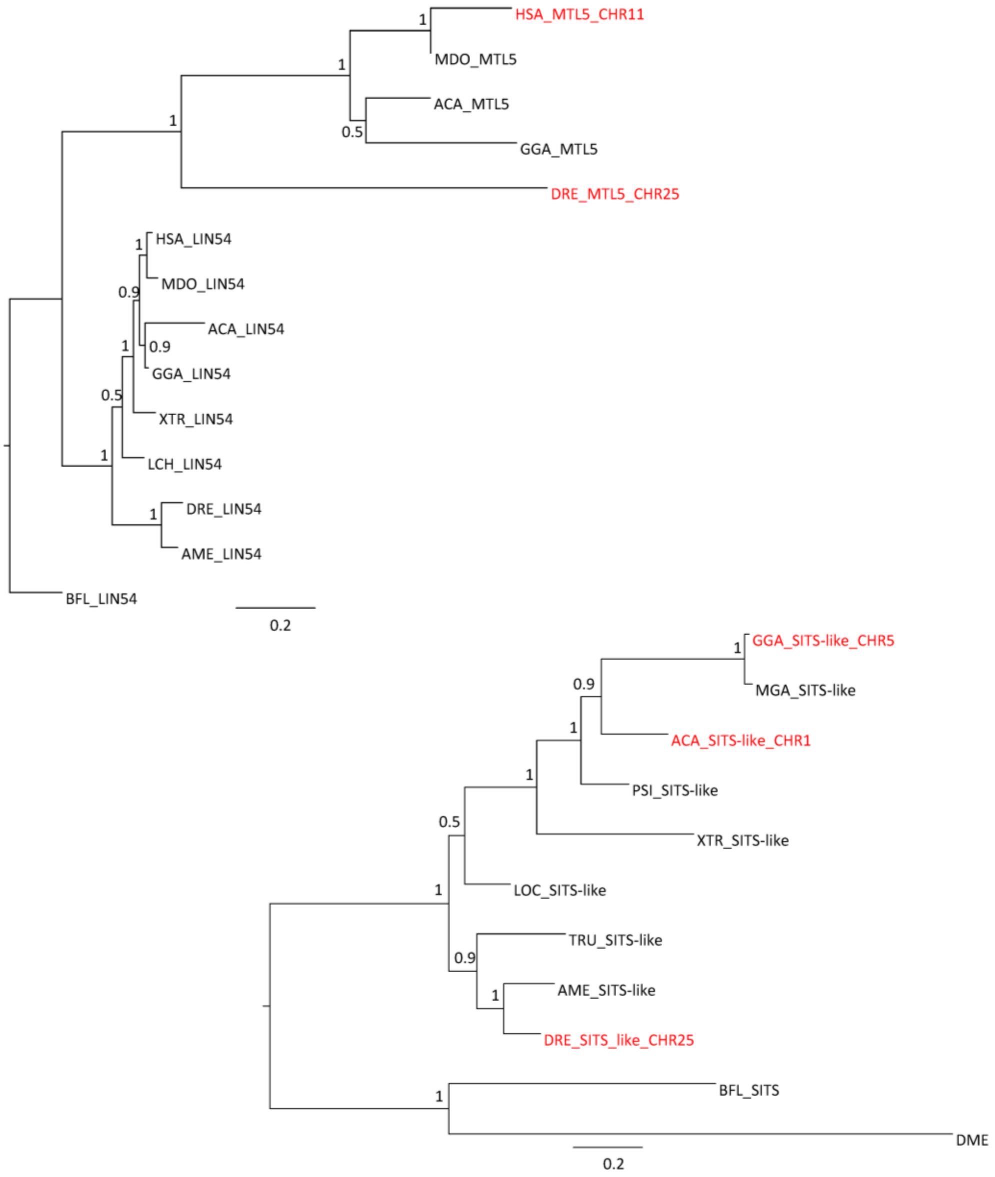
HSA – *H. sapiens*; SSC- *S. scrofa;* MMU – *M. musculus*; SHA – *S. harrisii*; MDO – *M. domestica*; ACA – *A. carolinensis* ; GGA – *G. gallus*; FPE – *F. peregrinus;* XTR – *X. tropicalis* ; LCH – *L. chalumnae* ; DRE – *D. rerio*; AME – *A. mexicanus;* TNI – *T. nigroviridis*. TFU – *T. fulvidraco;* GAC – *G. aculeatus* ; TRU – *T. rubripes* ; ONI – *O. niloticus* ; OLA – *O. latipes* ; CMI – *C. milii;* LER - *L. erinacea;* CIN – *C. instestinalis* ; DME – *D. melanogaster*



Phylogenetic analysis conducted in MEGA6 using Neighbor-Joining method . The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. The evolutionary distances were computed using the JTT matrix-based method and are in the units of the number of amino acid substitutions per site.

HSA – *H. sapiens*; SSC- *S. scrofa;* MMU – *M. musculus*; SHA – *S. harrisii*; MDO – *M. domestica*; ACA – *A. carolinensis* ; GGA – *G. gallus*; FPE – *F. peregrinus;* XTR – *X. tropicalis* ; LCH – *L. chalumnae* ; DRE – *D. rerio*; AME – *A. mexicanus;* TNI – *T. nigroviridis*. TFU – *T. fulvidraco;* GAC – *G. aculeatus* ; TRU – *T. rubripes* ; ONI – *O. niloticus* ; OLA – *O. latipes* ; CMI – *C. milii;* LER - *L. erinacea;* CIN – *C. instestinalis* ; DME – *D. melanogaster*

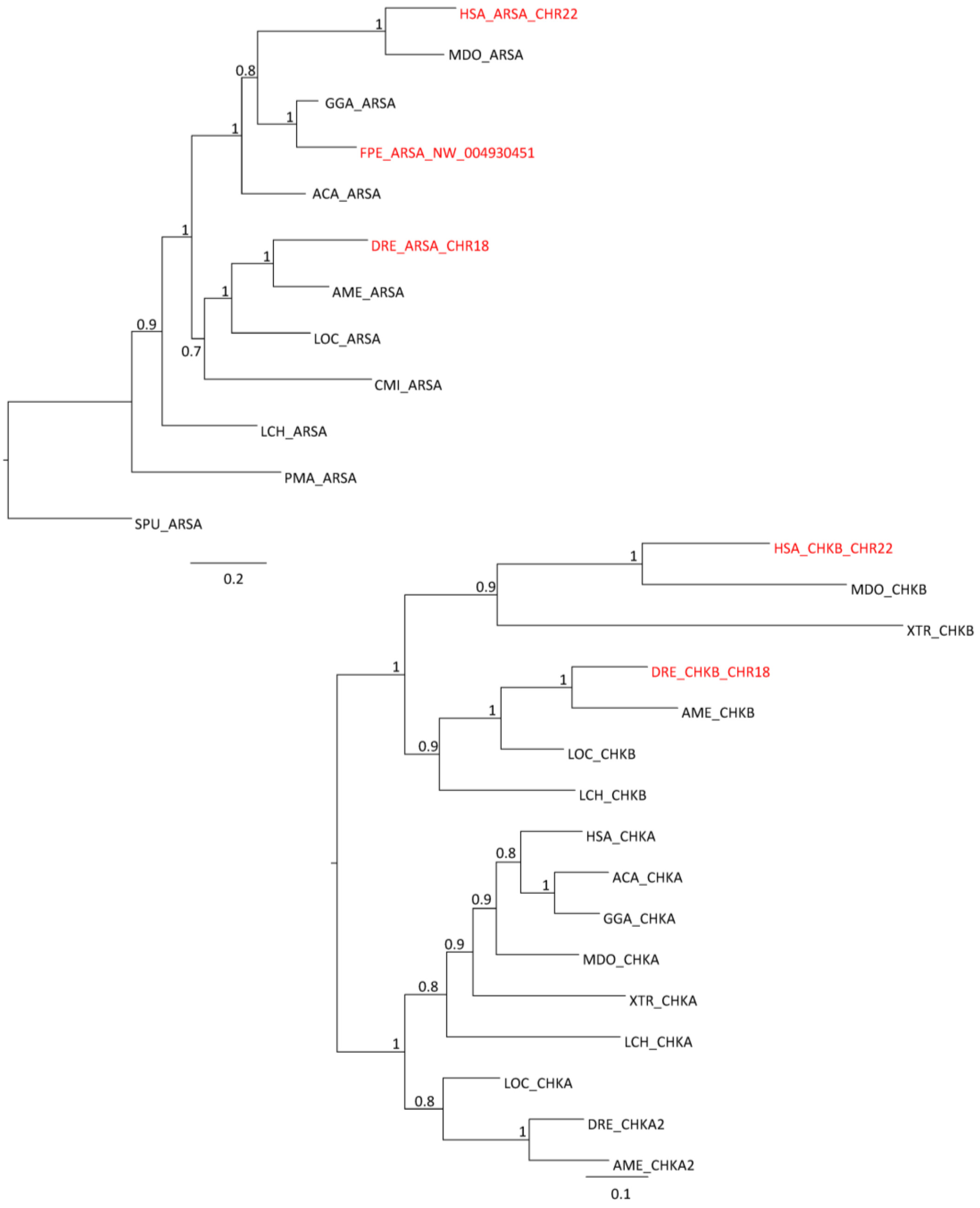
**Supporting information 4 Fig:** Supporting phylogenetic analysis of *Cpt1a* neighboring genes *Mlt5* and *Sits-like.*



Sequence alignment was performed using MAFFT with the L-INS-i method (14). The final alignment was curated in BioEdit version 7.2.5 (15) with the removal of all columns containing gaps. Molecular phylogenetic analysis by Maximum Likelihood was performed in PhyML with SMS (smart model selection) option, node values represent branch support using the aBayes algorithm. Genes neighbouring *Cpt1a* are depicted in red. with the corresponding chromosomal location.

HSA – *H. sapiens;* MMU – *M. musculus*; MDO – *M. domestica*; ACA – *A. carolinensis*; PSI – *P. sinensis;* GGA – *G. gallus;* MGA- *M. gallopavo* XTR – *X. tropicalis* ; LCH – *L. chalumnae*; DRE – *D. rerio*; AME – *A. mexicanus*; LOC- *L. oculatus*. BBE- *B. belcheri*. BFL-*B. floridae*

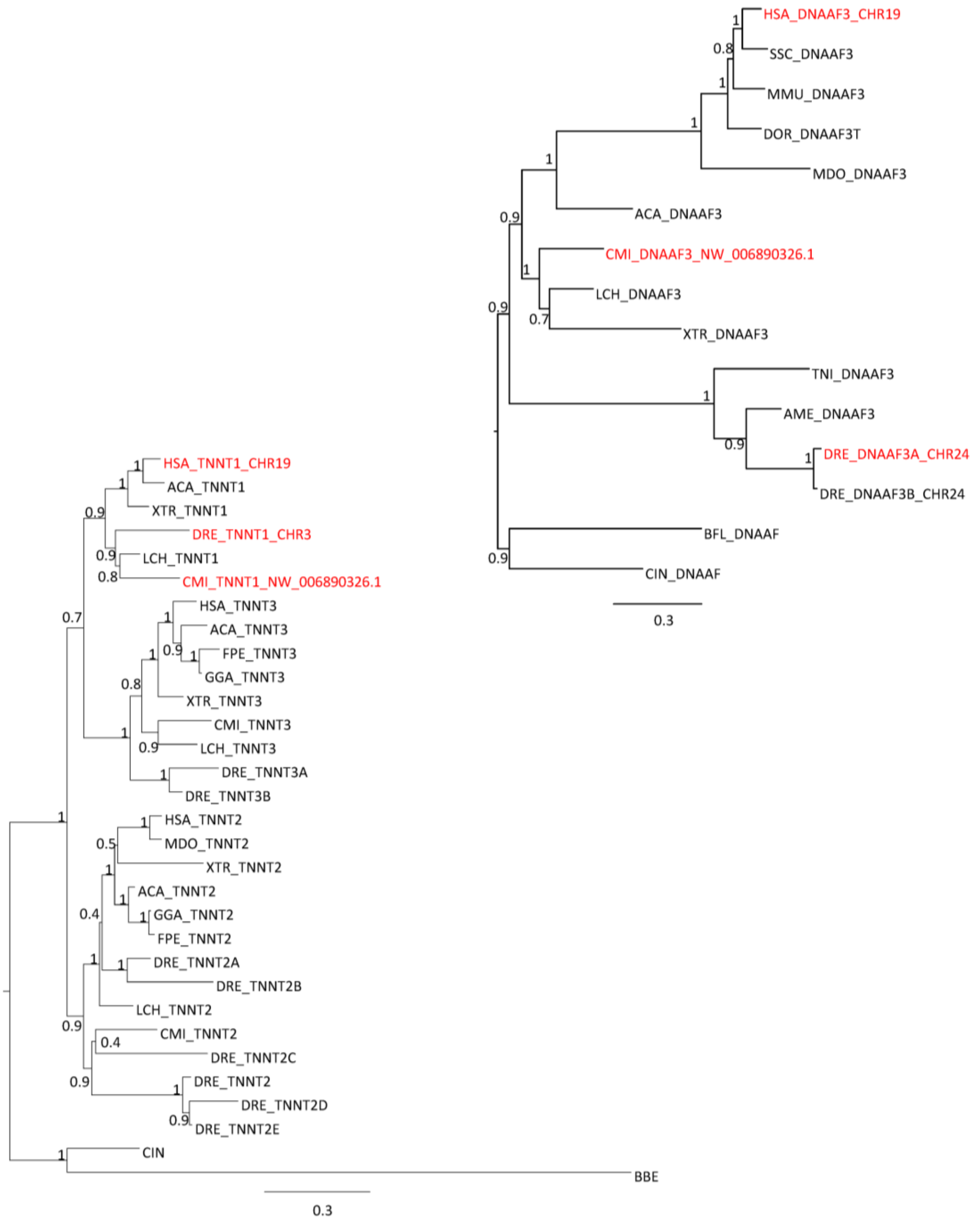
**Supporting information 5 Fig:** Supporting phylogenetic analysis of *Cpt1b* neighboring genes *Arsa* and *Chkb*.



Sequence alignment was performed using MAFFT with the L-INS-i method (14). The final alignment was curated in BioEdit version 7.2.5 (15) with the removal of all columns containing gaps. Molecular phylogenetic analysis by Maximum Likelihood was performed in PhyML with SMS (smart model selection) option, node values represent branch support using the aBayes algorithm. Genes neighbouring *Cpt1b* are depicted in red, with the corresponding chromosomal location.

HSA – *H. sapiens*; MDO – *M. domestica*; ACA – *A. carolinensis* ; GGA – *G. gallus*; FPE – *F. peregrinus;* XTR – *X. tropicalis* ; LCH – *L. chalumnae*; DRE – *D. rerio*; AME – *A. mexicanus;* LOC- *L. oculatus*, CMI – *C. milii,* PMA - *P. marinus*, SPU- *S. purpuratus* .

**Supporting information 6 Fig**: Supporting phylogenetic analysis of *Cpt1c* neighboring genes *Dnaaf3* and *Tnnt1*.



Sequence alignment was performed using MAFFT with the L-INS-i method (14). The final alignment was curated in BioEdit version 7.2.5 (15) with the removal of all columns containing gaps. Molecular phylogenetic analysis by Maximum Likelihood was performed in PhyML with SMS (smart model selection) option, node values represent branch support using the aBayes algorithm. Genes neighbouring *Cpt1c* are depicted in red, with the corresponding chromosomal location.

HSA – *H. sapiens*; SSC- *S. scrofa;* MMU – *M. musculus*; MDO – *M. domestica*; DOR – *D. ordii;* ACA – *A. carolinensis* ; GGA – *G. gallus*; FPE – *F. peregrinus;* XTR – *X. tropicalis* ; LCH – *L. chalumnae* ; CMI – *C. milii* ; DRE – *D. rerio*; AME – *A. mexicanus; TNI – T. nigroviridis*, CMI – *C. milii,,* CIN – *C. instestinalis;* BBE – *B. belcheri;* BFL – *B. floridae*

## V.2 Basal Gnathostomes Provide Unique Insights into the Evolution of Vitamin B12 Binders

Mó nica Lopes-Marques*, Raquel Ruivo*, Inês Delgado, Jonathan M. Wilson,

Neelakanteswar Aluru, L. Filipe C. Castro

(*Joint First authors)

**Genome Biology And Evolution - GBE**

(DOI: 10.1093/gbe/evu289)

# Basal Gnathostomes Provide Unique Insights into the Evolution of Vitamin B12 Binders

GBE

Mónica Lopes-Marques[1,2,†], Raquel Ruivo[1,†], Inês Delgado[1], Jonathan M. Wilson[1,3], Neelakanteswar Aluru[4], and L. Filipe C. Castro[1,5,*]

[1]CIIMAR—Interdisciplinary Centre of Marine and Environmental Research, CIMAR Associate Laboratory, UPorto—University of Porto, Portugal

[2]ICBAS—Institute of Biomedical Sciences Abel Salazar, UPorto—University of Porto, Portugal

[3]Department of Biology, Wilfred Laurier University-Waterloo, Ontario, Canada

[4]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts

[5]Department of Biology, Faculty of Sciences, UPorto—University of Porto, Portugal

*Corresponding author: E-mail: filipe.castro@ciimar.up.pt.

†These authors contributed equally to this work.

**Accepted:** December 23, 2014

**Data deposition:** Three sequences from genes isolated from *Leucoraja erinacea* have been deposited at NCBI GenBank under the following acession numbers: KP273226 (TCN2); KP273227 (TCN3); KP273228 (TCN1).

## Abstract

The uptake and transport of vitamin B12 (cobalamin; Cbl) in mammals involves a refined system with three evolutionarily related transporters: transcobalamin 1 (*Tcn1*), transcobalamin 2 (*Tcn2*), and the gastric intrinsic factor (*Gif*). Teleosts have a single documented binder with intermediate features to the human counterparts. Consequently, it has been proposed that the expansion of Cbl binders occurred after the separation of Actinopterygians. Here, we demonstrate that the diversification of this gene family took place earlier in gnathostome ancestry. Our data indicates the presence of single copy orthologs of the Sarcopterygii/Tetrapoda duplicates *Tcn1* and *Gif,* and *Tcn2*, in Chondrichthyes. In addition, a highly divergent Cbl binder was found in the Elasmobranchii. We unveil a complex scenario forged by genome, tandem duplications and lineage-specific gene loss. Our findings suggest that from an ancestral transporter, exhibiting large spectrum and high affinity binding, highly specific Cbl transporters emerged through gene duplication and mutations at the binding pocket.

**Key words:** cobalamin transport, genome duplications, gnathostomes.

## Background

Cobalamin (Cbl; Vitamin B12) is an essential nutrient for metazoans. It is required as the basis for two enzyme cofactors, methyl-Cbl and 5′-deoxyadenosyl-Cbl, for methionine synthase and methyl-malonyl-CoA mutase, respectively, involved in the folate and mutase pathways (Banerjee and Ragsdale 2003). Accordingly, Cbl is fundamental for the synthesis of nucleotides, branched-chain amino acids, and odd-chain fatty acids (Banerjee and Ragsdale 2003; Carmel et al. 2003). Animal diets must include Cbl because only microorganisms are able to synthesize this compound. In humans, Cbl deficiency leads, among others, to anemia and severe neurological dysfunction. Once ingested, an elaborate system involving protein binders is responsible for the absorption, transport, and cellular uptake. Mammalian species typically possess three Cbl binder genes: *Gif* (or gastric intrinsic factor), *Tcn1* (also known as haptocorrin), and *Tcn2* (also known as transcobalamin). This carrier diversity mirrors their specialization to different physiological environments and functional specificities. Interestingly, *Tcn1* is absent in some species such as mouse, rat, and probably the marsupial opossum (Greibe, Fedosov, Nexø, et al. 2012). In birds and amphibians two binders have been found, whereas in reptiles a similar gene repertoire to that of most mammals is found (Greibe, Fedosov, Nexø, et al. 2012). Teleosts such as zebrafish, trout, and salmon have a single documented binder (Greibe, Fedosov, Nexø, et al. 2012; Greibe, Fedosov, Sorensen, et al. 2012). This has led to the proposal that the evolutionary elaboration of Cbl binding proteins occurred "after" the divergence of Actinopterygians (Greibe, Fedosov, Nexø, et al. 2012). In effect, the zebrafish and trout Cbl binding protein exhibits mixed characteristics.

Structurally it resembles a hybrid of the full set of human Cbl binders. The sequence identity is closer to *Tcn2* (Greibe, Fedosov, Nexø, et al. 2012), the amino acid composition at the binding site is similar to *Tcn1* (Greibe, Fedosov, Nexø, et al. 2012; Greibe, Fedosov, Sorensen, et al. 2012), and it shows resistance toward degradation by trypsin comparable to *Gif* (Greibe, Fedosov, Nexø, et al. 2012; Greibe, Fedosov, Sorensen, et al. 2012). Consequently, it has been named *HIT* (an abbreviation for haptocorrin, intrinsic factor, and transcobalamin) denoting its intermediate and ancestral nature (Greibe, Fedosov, Nexø, et al. 2012).

To infer the exact evolutionary history of a gene family based on functional aspects, without considering phylogenetics and an adequate species sampling, can lead to inaccurate conclusions. Particularly relevant when addressing these issues is the role of gene/genome duplications and gene loss. For example, two rounds of whole-genome duplications (1R and 2R) have taken place in early vertebrate ancestry (Putnam et al. 2008). Additional events of whole-genome duplication have occurred, one in teleost ancestry (3R) (Jaillon et al. 2004), and a second specifically in salmonids (4R) (Berthelot et al. 2014). In this context, the repertoire of Cbl binding proteins in teleosts may represent a case of secondary lineage specific-gene loss after duplication and not an ancestral state.

Here, to distinguish between different evolutionary hypotheses, we analyzed the gene diversity and *loci* composition in a variety of vertebrate species, particularly in basal gnathostomes, Chondrichthyans.

## Materials and Methods

### Sequence Mining and Phylogenetic Analysis

*Tcn1*, *Gif*, and *Tcn2* sequences from all major vertebrate lineages and from the invertebrate species *Branchiostoma floridae* (amphioxus) and *Saccoglossus kowalevskii* (acorn worm) were identified in the Ensembl, GenBank, Skatebase (http://skatebase.org/) databases via tBLASTn and BLASTp searches using as reference annotated human Cbl binder sequences. Amino acid sequences were aligned with MAFFT alignment software (Katoh and Toh 2010) using default parameters and visualized and edited in Geneious v7.1.7. *Gallus gallus* TCN2 partial sequence (XP_427292.3) was excluded from the analysis. Although extensive searches were performed, we were unable to retrieve *Chelonia mydas Gif* and *Tcn1, Pelodiscus sinensis Tcn2* and *Gif*, as well as *Taeneopygia guttata Tcn2* possibly due to poor genome coverage. To infer the evolutionary model (LG + G) used for phylogenetic analysis, the alignment was stripped from columns containing gaps resulting in an alignment with 268 positions which was analyzed in Protest 3.3 (Abascal et al. 2009). Finally, phylogenetic analysis was performed on the online platform PhyML 3.0 (http://www.atgc-montpellier.fr/phyml/), and the aBayes algorithm was selected to calculate branch support (Guindon et al. 2010).

### Comparative Genomics

*Tcn1, Gif,* and *Tcn2* genes were localized onto the human chromosomes, the location of each gene and the neighboring genes were collected from Ensembl and GenBank databases. Gene *loci* in human were used as a reference to assemble the synteny maps of the remaining species. Gene families with multiple members (e.g., oxysterol binding protein - OSBP) flanking these genes in humans had their phylogenetic history determined to clarify if the duplication timing coincided with 2R (not shown).

### Gene Isolation and Expression Analysis

Adult *Leucoraja erinacea* were obtained from the Marine Biological Laboratory’s Marine Resources Center in Woods Hole, Massachusetts. Fish were collected from the coast of Woods Hole and maintained in 100 gallon recirculatory tanks under ambient conditions. All tissues were collected and preserved in RNAlater and stored at −20 °C. Total RNA was isolated using an Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, UK) according to the manufacturer’s recommendations, including the on-column treatment of isolated RNA with RNase-free DNase I. RNA concentration was calculated using Qubit fluorometer instrument (Invitrogen, Carlsbad CA), integrity confirmed by electrophoresis and the RNA stored at −80 °C until further use. Partial Tcn-like sequences were extended by Rapid amplification of cDNA ends (RACE) polymerase chain reaction (PCR) with the SMARTer 5′/3′ Kit (Clontech). *Leucoraja erinacea* full (or near full) open reading frames (ORFs) were obtained by PCR with the following primer sets: *Tcn1/Gif* primer forward 5′-GGGCAAGCAGTGGTATCAAC-3′, primer reverse 5′-GTTAGAGCGATGGGGAGAGG-3′; *Tcn3* primer forward 5′-ACGCAGAGTACATGGGGACT-3′, primer reverse 5′-TTATTAGTTGGCGGCGTTTC-3′ and *Tcn2* primer forward 5′-AGTGTCCACATTGCCTTGC-3′, primer reverse 5′-CCTGTAATTTGGGGCTTTCA-3. The cDNA was synthesized from 500 ng of total RNA with the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufacturer’s protocol. Tissue expression was determined through RT-PCR with intron flanking primers. PCR was performed using 2 μl of little skate cDNA and Phusion Flash high-fidelity Master Mix (FINNZYMES). PCR parameters were as follows: initial denaturation at 98 °C for 10 s, followed by 35 cycles of denaturation at 98 °C for 1 s, annealing for 5 s and elongation at 72 °C for 10 s, and a final step of elongation at 72 °C for 1 min. PCR products were then loaded onto 2% agarose gel stained with GelRed and run in TBE buffer at 80 V.

### Comparative Homology Modeling

*Tcn1/Gif*, *Tcn2*, and *Tcn3* amino acid sequences of *L. erinacea* were submitted to the online platform iTASSER (Zhang 2008; Roy et al. 2010) for modeling. The predicted structural models were analyzed and visualized using Open-Source PyMOL V1.3 (academic version; Schrödinger 2011).

## Results

We began by providing clarification of the orthology of the teleost single copy sequences. Our analysis shows that the teleost Cbl binder forms a monophyletic clade with Sarcopterygii *Tcn2* (fig. 1). Further evidence for the common origin of *Tcn2* sequences comes from synteny analysis (fig. 2). The *Tcn2 locus* is relatively well conserved in the examined species, with the exception of zebrafish (fig. 2). However, the holostean spotted gar *Tcn2 locus* (*Lepisosteus oculatus*), which diverged prior to the teleost-specific duplication (3R), retains synteny with tetrapods (e.g., *SLC35E4*; fig. 2). Thus, teleost sequences should be named *Tcn2* and not *HIT*. Additionally, we also find a novel *Tcn2* gene specific to salmon, probably resulting from the salmonid-specific genome duplication, 4R (fig. 1).

The branching pattern of our phylogenetic analysis provides additional clues on the probable timing of *Tcn1* and *Gif* emergence. Orthologs of both genes can be found in mammals (except *Tcn1* in rodents and the opossum) and reptiles. In contrast, amphibians and birds have a single gene but clearly grouping with other *Gif* sequences (fig. 1). In the basal Sarcopterygian coelacanth we found two *Gif/Tcn1-like* sequences. However, one is an incomplete sequence too short to include in the phylogenetic analysis (not shown; supplementary material, Supplementary Material online). The basal position of the complete coelacanth sequence in the tree suggests that *Gif* and *Tcn1* originated from a duplication event in the ancestor of Tetrapoda (fig. 1). However, the incomplete Cbl binder sequence of the coelacanth is flanked by genes whose orthologs in tetrapods localize to the *Gif* and *Tcn1* genomic location (fig. 2). Thus, both sequences might represent bona fide *Gif* and *Tcn1* orthologs. The completion of the partial coelacanth sequence as well as the investigation of the Cbl binder gene repertoire in lungfish should help to resolve this matter. Additionally, the syntenic composition of this *locus* in various lineages (fig. 2) in combination with the phylogenetic analysis supports the independent loss of *Tcn1* in amphibians, birds, and some mammalian species.

The overall evolutionary branching pattern suggests that the *Tcn2* and *Tcn1/Gif* gene duplication predates teleost radiation (fig.1), and so a *Gif/Tcn1* gene would have been independently lost in this lineage. Interestingly, additional *Tcn*-like sequences have been reported in teleosts (Greibe, Fedosov, Nexø, et al. 2012). These are shorter than typical TCN proteins, composed of a DUF4430 domain present in the C-terminus region, similarly to TCN proteins (supplementary material, Supplementary Material online). Although they could represent the remnants of an ancestral *Gif/Tcn1* gene(s), neither phylogenetics (not shown) nor synteny analysis (supplementary material, Supplementary Material online) clarifies their origin.

To further explore the evolutionary history of Cbl binders, we next investigated the gene repertoire in the most basal clade of jawed vertebrates, the Chondrichthyans. In the recent release of the elephant shark genome sequence, we identified an ortholog of *Tcn2* (fig. 1; Venkatesh et al. 2014). Furthermore, a second sequence was found in the transcriptome of the same species which branches basally to the *Tcn1* and *Gif* clade (fig.1), thus supporting an event of gene loss in the Actinopterygii lineage. We next examined the partial genome and transcriptome sequences of the little skate, *L. erinacea* (Wang et al. 2012). Surprisingly, we found three partial sequences with similarity to Cbl binders, which were further expanded by PCR to obtain full or near full-length sequences. Phylogenetic analysis indicates that two of these group with the *Tcn2* and *Tcn1/Gif* gene clades, respectively, as observed in the elephant shark (fig. 1), while the third represents an apparently novel Cbl gene lineage so far unique to little skate, which we name *Tcn3* (fig. 1). To envisage the evolutionary origin of this extra sequence without synteny data, which is currently unavailable for this species, is problematic. Nevertheless, paralogy analysis of the human *loci* containing *Tcn2* and *Tcn1/Gif* provides a plausible explanation (supplementary material, Supplementary Material online). In effect, these genes reside in genomic regions related by duplication dating back to 2R (Putnam et al. 2008). For example, the *OSBP* gene, which maps close to human *Tcn1/Gif* at chromosome 11 has a paralog, *OSBP2*, close to *Tcn2* at chromosome 22 (supplementary material, Supplementary Material online). Detailed analysis shows various gene families whose paralogs map in expected regions of paralogy (supplementary material, Supplementary Material online), thus indicating that *Tcn2* and the ancestor of *Tcn1* and *Gif* are 2R-generated paralogs. In this context, we put forward that *Tcn3* might represent a 2R paralog retained uniquely in Elasmobranchii (or Chondrichthyans) but subsequently lost in other gnathostome lineages, similarly to what has been described in other gene families (e.g., Mulley and Holland 2010; Hoffmann et al. 2011, 2012; Ravi et al. 2013). In an alternative scenario the Elasmobranchii *Tcn3* and *Tcn1/Gif* genes might represent true *Gif* and *Tcn1* orthologs respectively, whose phylogenetic relationships toward Sarcopterygii sequences have been obscured by sequence divergence. If so, the duplication of *Tcn1* and *Gif* would date back to the origin of gnathostomes. Interestingly, the expression of the so-called *Tcn3* in little skate is significantly higher in the stomach (supplementary material, Supplementary Material online), paralleling the mammalian *Gif*. Whether *Tcn3* is present in other Chondrichthyes species is also a pertinent question, in particular the elephant shark given its agastric condition (Castro et al. 2014). Further investigations, namely with the inclusion of synteny data should fully clarify the origin of *Tcn3*.

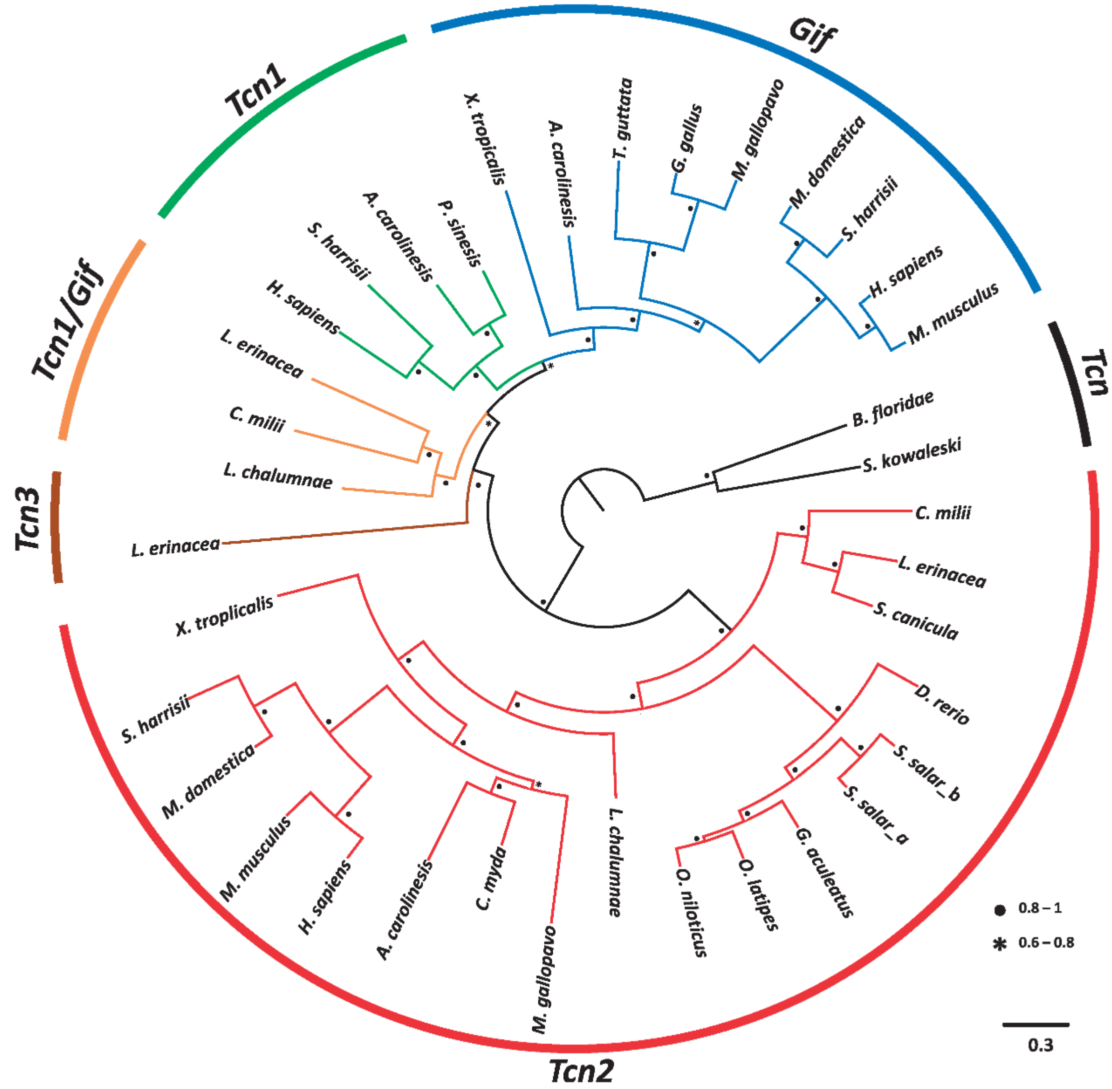


FIG. 1.—Maximum likelihood phylogenetic tree describing relationships among Cbl binding proteins from representative vertebrate *taxa* and two invertebrate deuterostomes. Node values represent branch support using the aBayes algorithm. Accession numbers for all sequences are provided in the supplementary material, Supplementary Material online.

## Discussion

In this study, we explored the evolutionary history of Cbl binding proteins in vertebrates. Our findings support a model where a single Cbl binder duplicated in early vertebrate ancestry as part of 2R (fig. 3). A later event of duplication (tandem) in the ancestor of either Sarcopterygii or Tetrapoda gave origin to *Gif* and *Tcn1*, with the latter being lost independently in amphibians, birds, and some mammalian species (fig. 3). In teleosts only *Tcn2* has been retained. We also found a novel, highly divergent Cbl binder in little skate, *Tcn3*, even though without synteny data we cannot firmly conclude on its evolutionary origin. Why exactly have different lineages retained such a variable repertoire of Cbl binders is difficult to establish *a priori*. Although all binders share a similar structure, they display distinct physiological functions. In mammals, the specificity toward Cbl is higher for GIF and TCN2 but substantially lower for TCN1 (Fedosov et al. 2007). In fact, gastric GIF and plasma TCN2 are required for Cbl absorption via receptor-mediated endocytosis in the ileum and target cells, respectively (Furger et al. 2012). TCN1, on the other hand, occurs in several body fluids, including saliva, milk, and plasma (Morkbak et al. 2007). When

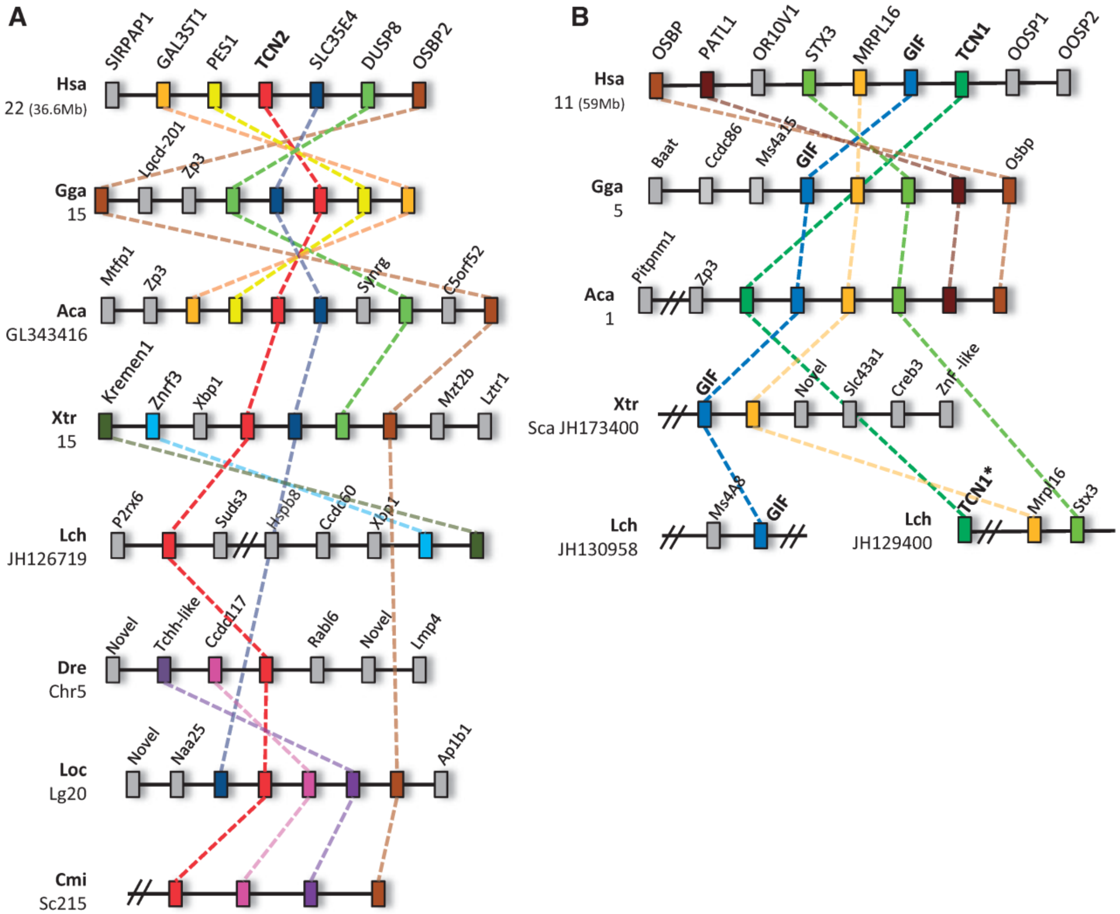


Fig. 2.—Synteny maps of *Tcn1*, *Gif*, and *Tcn2 loci*. (*A*) Detail of the *Tcn2 locus* which is highly conserved in major vertebrate lineages; (*B*) Detail of the *Tcn1* and *Gif locus*, depicting a highly conserved *locus* in tetrapods. The *Tcn1 locus* is disrupted in teleosts (not shown). Information is presently absent for the *Tcn1* and *Tcn3 loci* in the chondrichthyan lineage. Hsa, *Homo sapiens*; Gga, *Gallus gallus*; Aca, *Anolis carolinensis*; Xtr, *Xenopus tropicalis*; Lch, *Latimeria chalumnae*; Dre, *Danio rerio*; Loc, *Lepisosteus oculatus*; and Cmi, *Callorhinchus milii*. * denotes partial sequence. Double dashes denote gap.

compared with the other carriers, TCN1 exhibits higher binding affinity toward Cbl, faster binding kinetics, and lower specificity, binding other apparently inert corrinoids (Fedosov et al. 2007; Wuerges et al. 2007). In the lumen of the upper gut, TCN1 confers increased stability to the carrier-corrinoid complex at low pH conditions; yet, due to its reduced resistance to pancreatic proteolytic enzymes Cbl is relayed to GIF in the small intestine (Greibe, Fedosov, Sorensen, et al. 2012). Although TCN1 also binds the vast majority of Cbl in the plasma, its role there is more elusive. Given its glycosylation status (Wuerges et al. 2007; Furger et al. 2012), TCN1-dependent corrinoid uptake from plasma via liver asialoglycoprotein receptors was suggested (Furger et al. 2012). Thus, TCN1 likely recycles Cbl, to an additional round of intestinal absorption, as well as acts as a scavenger in the blood for toxic Cbl-derived molecules, leading to its excretion (Wuerges et al. 2007). Interestingly, this later capacity is not unique to TCN1. The lack of TCN1 in mice is apparently functionally compensated by the action of TCN2 (Hygum et al. 2011). For example, the murine TCN2 is capable of binding to a Cbl analogue, cobinamide, just like the human TCN1 transporter (Hygum et al. 2011). Similarly, in zebrafish, TCN2 has been found to bind the analogue cobinamide, while still efficiently binding Cbl (Greibe, Fedosov, Nexø, et al. 2012). This functional plasticity is apparently structurally determined (table 1). Despite poor sequence conservation Cbl carriers retain a similar two domain-structure, α and β, that clamp corrinoids (Wuerges et al. 2007), also visible in the three little skate carriers (supplementary material, Supplementary Material online). In humans, several features apparently sustain their differential affinity and selectivity: interdomain contacts and complementarity and carrier-specific ligand interactions (Wuerges et al.

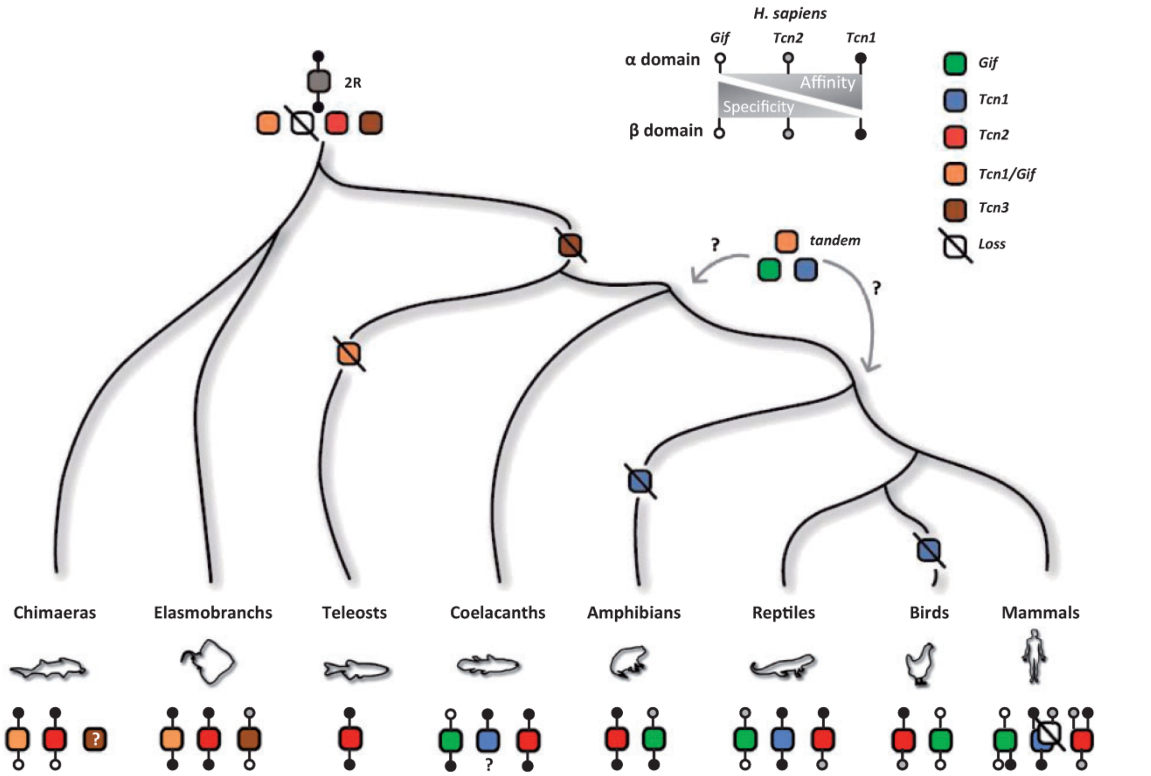


**Fig. 3.**—Evolutionary model of Cbl binding proteins in vertebrates. Specificity and affinity gradients illustrate the binding properties of the human carriers (top). Grayscale circles indicate α and β domain signature motif conservation in vertebrate carriers deduced from table 1 (bottom).

2007; Furger et al. 2013) (table 1; supplementary material, Supplementary Material online). However, some reported structural differences are human specific and cannot be transposed to other groups (e.g. an additional disulfide bridge in human TCN1). The lower specificity and higher affinity of human TCN1 were justified by the presence of two amino acid clusters (table 1; supplementary material, Supplementary Material online; Wuerges et al. 2007; Furger et al. 2013). On the β-domain three bulky residues (Arg$^{357}$, Trp$^{359}$, and Tyr$^{362}$), the first human specific, provide hydrophobic contacts and stabilize the TCN1-corrinoid complex, compensating for the lack of the nucleotide moiety in cobinamide, a baseless corrinoid (Wuerges et al. 2007). Human GIF and TCN2 exhibit single bulky residues at distinct positions, Trp$^{359}$ or Tyr$^{362}$, respectively. The second motif, on the α-domain (TNYYQ), was suggested to form four H-bonds with the central corrinoid ring; a decreasing number of possible H-bonds is observed in TCN2 and GIF, corroborating the decline in the thermal stability of the carrier-corrinoid complexes and the gradual decrease of affinity toward Cbl (TCN1>TCN2>GIF) (Furger et al. 2013). Thus, TCN1 has been suggested to act as a scavenger in the blood for toxic Cbl-derived molecules (Wuerges et al. 2007). The occurrence of these motifs seems highly plastic throughout gnathostome evolution; yet scavenger-like carriers, retaining one or both amino acid motif signatures, are observed in all the examined species, including in the preduplicated TCN of cephalochordates (table 1; supplementary material, Supplementary Material online). For instance, in mouse, TCN2 retains the TNYYQ binding motif, but not the pair of bulky residues (table 1). Nonetheless, this carrier functionally behaves like TCN1: exhibiting an affinity toward Cbl comparable to that of human TCN1 and ability to bind cobinamide, yet with lower efficiency (Hygum et al. 2011). Similar results were obtained with teleost TCN2 (Greibe, Fedosov, Nexø, et al. 2012), which appears to display the full set of motifs, with the Tyr residue of the β-domain replaced by a similarly bulky Phe (table 1). Marsupials and birds, on the other hand, lack a true TNYYQ-like carrier but retain the second motif of aromatic residues (table 1). Thus, it is plausible to hypothesize that these carriers are able to bind and stabilize corrinoids, other than Cbl, to some extent. Although ligand recognition alone does not fully account for the variable number of Cbl carriers in the examined species, it illustrates the plasticity of these

**Table 1**
Cross-Species Variation of α and β Cbl Domain Signature Motifs

| | *GIF* | | *TCN2* | | *TCN1* | |
|---|---|---|---|---|---|---|
| | α<br>$[S/T]^1XXXX$ | β<br>ΩXXX | α<br>$[T/S]^1[S/T]YYQ^1$ | β<br>XXXΩ | α<br>$[T/S]^1N^2YYQ^1$ | β<br>ΩXXΩ |
| *Hsa GIF* | ✓ | ✓ | | | | |
| *Mmu GIF* | ✓ | ✓ | | | | |
| *Sha GIF* | ✓ | | | | | ✓ |
| *Mdo GIF* | ✓ | | | | | ✓ |
| *Aca GIF* | | ✓ | ✓ | | | |
| *Mga GIF* | ✓ | ✓ǂ | | | | |
| *Xtr GIF* | | | ✓ | | | ✓ |
| *Lch GIF* | ✓ | | | | | ✓ |
| *Hsa TCN1* | | | | | ✓ | ✓ |
| *Sha TCN1* | | | ✓ | | | ✓ |
| *Aca TCN1* | | | | | ✓ | ✓ |
| *Lch TCN1 (partial)* | | ? | | ? | ✓ | ? |
| *Cmi TCN1/GIF* | | ✓ | | | ✓ | |
| *Ler TCN1/GIF* | | | | | ✓ | ✓ |
| *Hsa TCN2* | | | ✓ | ✓ | | |
| *Mmu TCN2* | | | | ✓ | ✓ | |
| *Sha TCN2* | | | ✓ | ✓ | | |
| *Mdo TCN2* | | | ✓ | ✓ | | |
| *Aca TCN2* | | | | ✓ | ✓ | |
| *Mga TCN2* | | | ✓ | | | ✓ |
| *Xtr TCN2* | | | | | ✓ | ✓ |
| *Lch TCN2* | | | | | ✓ | ✓ |
| *Dre TCN2* | | | | | ✓ | ✓ |
| *Gac TCN2* | | | | | ✓ | ✓ |
| *Ssa TCN2* | | | | | ✓ | ✓ |
| *Cmi TCN2* | | ✓ | | | ✓ | |
| *Ler TCN2* | | | | | ✓ | ✓ |
| *Ler TCN3* | | ✓§ | ✓ | | | |
| *Bfl TCN2/TCN1/GIF* | | | | | ✓# | ✓ |
| *Sko TCN2/TCN1/GIF* | | | | | ✓# | ✓ |

NOTE.—The number of H-bonds formed between the α–domain of the human carriers and the corrinoid ring are indicated in superscript ($^{1/2}$). Ω represents the bulky hydrophobic residues of the β-domain. Gradual shifts in human carrier affinity and specificity are represented in the diagram above. §, ΩXΩX; ǂ, XXXX; #, NNXXQ; ?, unkown. Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Sha, *Sarcophilus harrisii*; Mdo, *Monodelphis domestica*; Aca, *Anolis carolinensis*; Mga, *Meleagris gallopavo*; Xtr, *Xenopus tropicalis*; Lch, *Latimeria chalumnae*; Cmi, *Callorhinchus milii*; Ler, *Leucoraja erinacea*; Dre, *Danio rerio*; Gac, *Gasterosteus aculeatus*; Ssa, *Salmo salar*; Bfl, *Branchiostoma floridae*; and Sko, *Saccoglossus kowalevskii*.

proteins (fig. 3). Overall, we suggest that from an ancestral protein with high affinity but low specificity toward Cbl, the increase in binding specificity was acquired (and lost) through gene duplication and recurrent mutations at the carrier Cbl binding pocket (fig. 3 and table 1). Conversely, carriers exhibiting large spectrum and high affinity binding seem persistent, as are some binders with mixed profiles (fig. 3). The ancestral condition, retained in teleost and Elasmobranchii TCN2, shifted to additional carriers upon duplication events, as seen in most mammalian and reptile TCN1. In agreement, mouse TCN2 recapitulates the ancestral phenotype upon TCN1 loss. This functional shift possibly paralleled the acquisition of novel features in mammalian Cbl metabolism, notably membrane receptor recognition in target cells (Quadros et al. 2009). Our findings illustrate the decisive importance of basal gnathostomes to clarify gene family evolution and physiological diversity.

## Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

## Acknowledgments

The authors thank two anonymous reviewers for their helpful comments. This work was funded by the Fundação para a Ciência e a Tecnologia (FCT) project PEst-C/MAR/LA0015/2013. PhD grant SFRH/BD/84238/2012 awarded to M.L-.M. Postdoctoral grant SFRH/BPD/72519/2010 awarded to R.R.

## Literature Cited

Abascal F, Zardoya R, Posada D. 2009. ProtTest:selection of best-fit models of protein evolution. Bioinformatics 21(9):2104–2105.

Banerjee R, Ragsdale SW. 2003. The many faces of vitamin B12: catalysis by cobalamin-dependent enzymes. Annu Rev Biochem. 72: 209–247.

Berthelot C, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 5:3657.

Carmel R, Melnyk S, James SJ. 2003. Cobalamin deficiency with and without neurologic abnormalities: differences in homocysteine and methionine metabolism. Blood 101(8):3302–3308.

Castro LF, et al. 2014. Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history. Proc Biol Sci. 281: 20132669.

Fedosov SN, Fedosova N, Kräutler B, Nexø E, Petersen TE. 2007. Mechanisms of discrimination between cobalamins and their natural analogues during their binding to the specific B12-transporting proteins. Biochemistry 46(21):6446–6458.

Furger E, et al. 2012. Comparison of recombinant human haptocorrin expressed in human embryonic kidney cells and native haptocorrin. PLoS One 7(5):e37421.

Furger E, Frei DC, Schibli R, Fischer E, Prota AE. 2013. Structural basis for universal corrinoid recognition by the cobalamin transport protein haptocorrin. J Biol Chem. 288(35):25466–25476.

Greibe E, Fedosov S, Nexø E. 2012. The cobalamin-binding protein in zebrafish is an intermediate between the three cobalamin-binding proteins in human. PLoS One 7(4):e35660.

Greibe E, Fedosov S, Sorensen B, et al. 2012. A single rainbow trout cobalamin-binding protein stands in for three human binders. J Biol Chem. 287(40):33917–33925.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321[cited 2014 July 4].

Hoffmann FG, Opazo JC, Storz JF. 2011. Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. Genome Biol Evol. 3:588–600.

Hoffmann FG, Opazo JC, Storz JF. 2012. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Mol Biol Evol. 29(1):303–312.

Hygum K, et al. 2011. Mouse transcobalamin has features resembling both human transcobalamin and haptocorrin. PLoS One 6(5):e20638.

Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431: 946–957.

Katoh K, Toh H. 2010. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26(15):1899–1900.

Morkbak AL, Poulsen SS, Nexø E. 2007. Haptocorrin in humans. Clin Chem Lab Med. 45(12):1751–1759.

Mulley JF, Holland PWH. 2010. Parallel retention of *Pdx2* genes in cartilaginous fish and coelacanths. Mol Biol Evol. 27(10): 2386–2391.

Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198):1064–1071.

Quadros EV, Nakayama Y, Sequeira JM. 2009. The protein and the gene encoding the receptor for the cellular uptake of transcobalamin-bound cobalamin. Blood 113(1):186–192.

Ravi V, et al. 2013. Sequencing of *Pax6* loci from the elephant shark reveals a family of Pax6 genes in vertebrate genomes, forged by ancient duplications and divergences. PLoS Genet. 9(1): e1003177.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 5: 725–738.

Schrödinger L. 2011., The PyMOL molecular graphics system version 1.3. Schrödinger, LLC.

Venkatesh B, et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505(7482):174–179.

Wang Q, et al. 2012. Community annotation and bioinformatics workforce development in concert-Little Skate Genome Annotation Workshops and Jamborees. Database 2012:bar064 [cited 2015 December 5].

Wuerges J, Geremia S, Randaccio L. 2007. Structural study on ligand specificity of human vitamin B12 transporters. Biochem J. 403(3): 431–440.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40[cited 2014 June 5].

**Associate editor:** B. Venkatesh

## SUPPLEMENTARY MATERIAL

**Supplementary table 1**. List of sequences used for phylogenetic analysis.

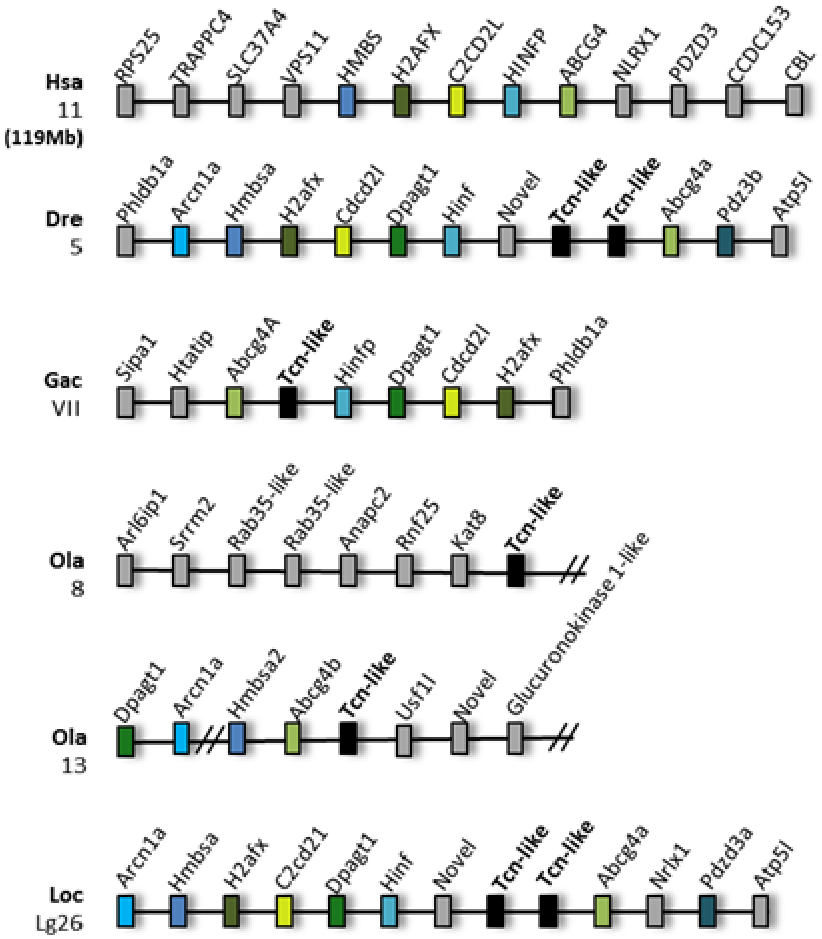
| Species | Gene | Accession number | Species | Gene | Accession number |
|---|---|---|---|---|---|
| ***Anolis carolinensis*** | Tcn1 | XP_003215131.1 | ***Monodelphis domestica*** | Tcn2 | XP_001380461.1 |
| | Tcn2 | XP_008117996.1 | | Gif | XP_007497607.1 |
| | Gif | XP_008108890.1 | ***Meleagris gallopavo*** | Tcn2 | ENSMGAP00000015262 |
| ***Branchiostoma floridae*** | Tcn | XP_002605724.1 | | Gif | ENSMGAP00000003005 |
| ***Callorhinchus milii*** | Tcn1/Gif | JW870881.1 | ***Mus musculus*** | Tcn2 | NP_056564.1 |
| | Tcn2 | NP_001279906.1 | | Gif | NP_032144.2 |
| ***Chelonia mydas*** | Tcn2 | EMP27579.1 | ***Oryzias latipes*** | Tcn2 | XP_004072417.1 |
| ***Danio rerio*** | Tcn2 | NP_001116703.1 | ***Oreochromis niloticus*** | Tcn2 | XP_005473430.1 |
| ***Gasterosteus aculeatus*** | Tcn2 | ENSGACP00000011296 | ***Pelodiscus sinesis*** | Tcn1 | XP_006110368.1 |
| ***Gallus gallus*** | Gif | XP_001233885.2 | ***Scyliorhinus canicula*** | Tcn2 | SSC-transcript-ctg15613 |
| ***Homo sapiens*** | Tcn1 | NP_001053.2 | ***Sarcophilus harrisii*** | Tcn1 | XP_003774045.1 |
| | Tcn2 | NP_000346.2 | | Tcn2 | XP_003762764.1 |
| | Gif | NP_005133.2 | | Gif | XP_003774041.1 |
| ***Latimeria chalumnae*** | Gif | XP_006013605.1 | ***Salmo salar*** | Tcn2a | NP_001133733.1 |
| | Tcn1 partial | XP_006011707.1 | | Tcn2b | ACN10392.1 |
| | Tcn2 | XP_005993347.1 | ***Saccoglossus kowaleski*** | Tcn | XP_002734140.1 |
| ***Leucoraja erinacea*** | Tcn1/Gif | KP273228 | ***Taeniopygia guttata*** | Gif | XP_002196005.2 |
| | Tcn2 | KP273226 | ***Xenopus tropicalis*** | Tcn2 | NP_001184035.1 |
| | Tcn3 | KP273227 | | Gif | XP_002941420.1 |

## Supplementary figures

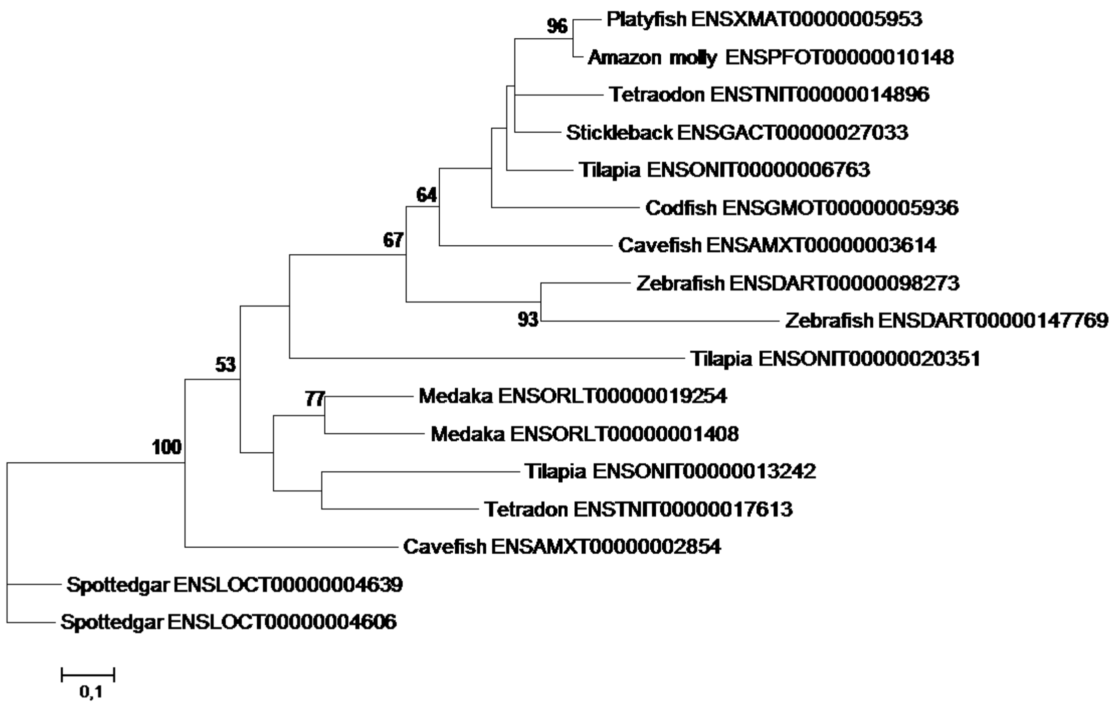
```
Hs_TCN1_NP_001053.2            1   MRQS------ ---------- HQLPLVGLLL FSF--I-PSQ LCEICEVSEE NYIRLKPL-- ---------- ---LNTMIQS  46
Hs_GIF_NP_005133.2             1   -MAW------ ---------- FALYLLSLLW ATA--GTSTQ TQSSCSVPSA QEPLVNGI-- ---------- ---QVLMENS  46
Hs_TCN2_NP_000346.2            1   ---------- ---------- --MRHLGAFL FLL--GVLGA LTEMCEIPEM DSHLVEKLGQ HLLPWMDRLS LEHLNPSIYV  56
Medaka_ENSORLT00000019254      1   ---------- ---------- --MKEPALIA AAL--L---- ---------- ---------- ---------- ---LLLPAAL  19
Medaka_ENSORLT00000001408      1   MAFR------ ---------- VIFTAAVLIL LTQ--AGDEG DDEGANSDRG SHAGAAAL-- ---------- ---LRHSDTV  47
Tilapia_ENSONIT00000013242     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Tetradon_ENSTNIT00000017613    1   ---------- ---------- ---MKLVLLS AAL--L---- ---------- ---LLLPA-- ---------- ---ARPERHQ  23
Cavefish_ENSAMXT00000002854    1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---MPPTAFT  7
Stickleback_ENSGACT00000027033 1   ---------- ---------- MAPRTAALLS AGF--L---- ---------- ---LL----- ---------- ---LTQRALT  23
Platyfish_ENSXMAT00000005953   1   MRAWSSINTD VKPAGETSHR MAPRLAAQLS VSFLLL---- ---------- ---LL----- ---------- ---LPHEAIT  45
Amazonmolly_ENSPFOT00000010148 1   MRGWSSINTD VKPGGETSHR MAPRPAALLS VSF-LL---- ---------- ---LL----- ---------- ---LPHEAIA  44
Tilapia_ENSONIT00000006763     1   ---------- ---------- MALRTSALLS VGF--F---- ---------- ---LL----- ---------- ---LTCGALT  23
Tetraodon_ENSTNIT00000014896   1   ---------- ---------- MRLRT-PLLS VGV--L---- ---------- ---LL----- ---------- ---LTGGVLT  22
Codfish_ENSGMOT00000005936     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Cavefish_ENSAMXT00000003614    1   ---------- ---------- LALSLVIFLC STV--L---- ---------- ---ICEPA-- ---------- ---LHIRAIQ  26
Zebrafish_ENSDART00000098273   1   ---------- ---------- MALTAISLLC FTA--L---- ---------- ---LCFPA-- ---------- ---LGLPADS  26
Zebrafish_ENSDART00000147769   1   ---------- ---------- MTLTAITLLC FAA--L---- ---------- ---LPFPG-- ---------- ---LGKGGHS  26
Spottedgar_ENSLOCT00000004606  1   ---------- ---------- MAL-AVTMIL STV--L---- ---------- ---LLVPA-- ---------- ---LLVQPES  25
Spottedgar_ENSLOCT00000004639  1   ---------- ---------- MAL-AVTMIL STV--L---- ---------- ---LLVPA-- ---------- ---LLVQAES  25
Tilapia_ENSONIT00000020351     1   ---------- ---------- --MKKPALLS AVL------- ---------- ---LLLFF-- ---------- ---VGTSAQG  23

Hs_TCN1_NP_001053.2            47  NYNRGTSAVN VVLSLKLVGI QIQTLMQKMI QQIKYNVKSR LSDVSSGELA LIILALGVCR NAEENLIYDY HLIDKLENKF  126
Hs_GIF_NP_005133.2             47  VTSSAYPNPS ILIAMNLAGA YNLKAQKLLT YQLMSSDNND ---LTIGQLG LTIMALTSSC ---------R DPGDKVSILQ  114
Hs_TCN2_NP_000346.2            57  GLRLSSLQAG TKEDLYLHSL KLGYQQCLLG SAFSEDDGDC QGKPSMGQLA LYLLALRANC EFVRGHKGDR LVSQLKWFLE  136
Medaka_ENSORLT00000019254      20  TQNF------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  23
Medaka_ENSORLT00000001408      48  RQDN------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  51
Tilapia_ENSONIT00000013242     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Tetradon_ENSTNIT00000017613    24  RGSV------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Cavefish_ENSAMXT00000002854    8   SDVE------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  11
Stickleback_ENSGACT00000027033 24  ETGP------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Platyfish_ENSXMAT00000005953   46  NQGL------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  49
Amazonmolly_ENSPFOT00000010148 45  NQGL------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  48
Tilapia_ENSONIT00000006763     24  DTDE------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Tetraodon_ENSTNIT00000014896   23  NAGP------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  26
Codfish_ENSGMOT00000005936     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Cavefish_ENSAMXT00000003614    27  TEEL------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Zebrafish_ENSDART00000098273   27  GKLE------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Zebrafish_ENSDART00000147769   27  GEQH------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Spottedgar_ENSLOCT00000004606  26  S--------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  26
Spottedgar_ENSLOCT00000004639  26  SEAR------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  29
Tilapia_ENSONIT00000020351     24  NTNF------ ---------- ---------- ---------- ---------- ---------- ---------- ----------  27

Hs_TCN1_NP_001053.2            127 QAEIENMEAH NGTPLTNYYQ LSLDVLALCL FNGNYSTAEV VNHFTPENKN YYFGSQFSVD TGAMAVLALT CVKKSLINGQ  206
Hs_GIF_NP_005133.2             115 RQMENWAPSS PNAEASAFYG PSLAILALCQ KNSEATLPIA VRFAKTLLAN ---SSPFNVD TGAMATLALT CMYNKI---P  188
Hs_TCN2_NP_000346.2            137 DEKRAIGHDH KGHPHTSYYQ YGLGILALCL HQKRVHDSVV DKLLYAVEPF H--QGHHSVD TAAMAGLAFT CLKRS-----  209
Medaka_ENSORLT00000019254      23  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  23
Medaka_ENSORLT00000001408      51  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  51
Tilapia_ENSONIT00000013242     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Tetradon_ENSTNIT00000017613    27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Cavefish_ENSAMXT00000002854    11  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  11
Stickleback_ENSGACT00000027033 27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Platyfish_ENSXMAT00000005953   49  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  49
Amazonmolly_ENSPFOT00000010148 48  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  48
Tilapia_ENSONIT00000006763     27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Tetraodon_ENSTNIT00000014896   26  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  26
Codfish_ENSGMOT00000005936     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Cavefish_ENSAMXT00000003614    30  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Zebrafish_ENSDART00000098273   30  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Zebrafish_ENSDART00000147769   30  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Spottedgar_ENSLOCT00000004606  26  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  26
Spottedgar_ENSLOCT00000004639  29  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  29
Tilapia_ENSONIT00000020351     27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27

Hs_TCN1_NP_001053.2            207 IKADEGSLKN ISIYTKSLVE KILSEKKENG LIGNTFSTGE AMQALFVSSD YYNENDWNCQ QTLNTVLTEI SQGAFSNPNA  286
Hs_GIF_NP_005133.2             189 VGSEEGYRSL FGQVLKDIVE KISMKIKDNG IIGDIYSTGL AMQALSVTPE -PSKKEWNCK KTTDMILNEI KQGKFHNPMS  267
Hs_TCN2_NP_000346.2            209 -NFNPGRRQR ITMAIRTVRE EILKAQTPEG HFGNVYSTPL ALQFLMTSPM RGAELGTACL KARVALLASL QDGAFQNALM  288
Medaka_ENSORLT00000019254      23  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  23
Medaka_ENSORLT00000001408      51  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  51
Tilapia_ENSONIT00000013242     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Tetradon_ENSTNIT00000017613    27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Cavefish_ENSAMXT00000002854    11  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  11
Stickleback_ENSGACT00000027033 27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Platyfish_ENSXMAT00000005953   49  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  49
Amazonmolly_ENSPFOT00000010148 48  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  48
Tilapia_ENSONIT00000006763     27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
Tetraodon_ENSTNIT00000014896   26  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  26
Codfish_ENSGMOT00000005936     1   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  1
Cavefish_ENSAMXT00000003614    30  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Zebrafish_ENSDART00000098273   30  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Zebrafish_ENSDART00000147769   30  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  30
Spottedgar_ENSLOCT00000004606  26  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  26
Spottedgar_ENSLOCT00000004639  29  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  29
Tilapia_ENSONIT00000020351     27  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  27
```

```
Hs_TCN1_NP_001053.2            287 AAQVLPALMG KTFLDINKDS SCVSASGNFN ISADEPITVT PPDSQSYISV NYSVRINE-- ------TYFT NVTVLNGSVF  358
Hs_GIF_NP_005133.2             268 IAQILPSLKG KTYLDVPQVT CSPDHEVQPT LPSNPGPG-- -PTSASNITV IYTIN-NQLR GVELLFNETI NVSVKSGSVL  343
Hs_TCN2_NP_000346.2            289 ISQLLPVLNH KTYIDLIFPD CLAPRVML-- ----EPAAET IPQTQEIISV TLQV----LS LLP---PYRQ SISVLAGSTV  355
Medaka_ENSORLT00000019254      23  ---------- ---------- ---------- ---------- -----DPAPI QIVVK-NSFL EEE---PLAF NSHVAHRGIL  54
Medaka_ENSORLT00000001408      51  ---------- ---------- ---------- ---------- -----NLDPI TIVVK-NKFQ GV----KKTY NASVAYRGIL  81
Tilapia_ENSONIT00000013242     1   ---------- ---------- ---------- ---------- ------MIPI AIMVK-NTLQ NKP---LQTY KTEVISGGIL  30
Tetradon_ENSTNIT00000017613    27  ---------- ---------- ---------- ---------- ------SVPI AVVVQ-NLLH NKP---SLTF TTSTADGGIL  57
Cavefish_ENSAMXT00000002854    11  ---------- ---------- ---------- ---------- -----SPYKI SLVVY-NSLT TAK---NLTF STDIAYRGIL  42
Stickleback_ENSGACT00000027033 27  ---------- ---------- ---------- ---------- -----GARPI RLSVE-NDLS NIT---PESY FSSVVEGGVL  58
Platyfish_ENSXMAT00000005953   49  ---------- ---------- ---------- ---------- -----ESLPI RLTVE-NDLH NMA---PESF SSTVVKEGVL  80
Amazonmolly_ENSPFOT00000010148 48  ---------- ---------- ---------- ---------- -----TSLPI RLTVE-NDLS NTA---PESF SSSVVEGGVL  79
Tilapia_ENSONIT00000006763     27  ---------- ---------- ---------- ---------- -----GSLSI KLSVE-NELS NEP---LKSY SSSVVEGGVL  58
Tetraodon_ENSTNIT00000014896   26  ---------- ---------- ---------- ---------- -----AALPL RLSVV-NTLS DMV---PGSY SSSVVEGGVL  57
Codfish_ENSGMOT00000005936     1   ---------- ---------- ---------- ---------- --------PI RVSVEGRGLS SEA---TGSY SGSVVEGGVL  29
Cavefish_ENSAMXT00000003614    30  ---------- ---------- ---------- ---------- -----KPVPI RVTVK-DEFS AS----SSFF QTSVLEGGVL  60
Zebrafish_ENSDART00000098273   30  ---------- ---------- ---------- ---------- ------EIPV KVTIV-NDFT NE----QLSY STTVIQEGLM  59
Zebrafish_ENSDART00000147769   30  ---------- ---------- ---------- ---------- -GGVPGQVSI NVVVT-NKFA NE----LNTY PVTAPKGMPI  64
Spottedgar_ENSLOCT00000004606  26  ---------- ---------- ---------- ---------- -----GWSPI LLSVR-NAID QKA---PLSF RGSVPYRGSL  57
Spottedgar_ENSLOCT00000004639  29  ---------- ---------- ---------- ---------- ----SKWSPI QLSVE-NAIE STP---PLIF KGSVPYRGVL  61
Tilapia_ENSONIT00000020351     27  ---------- ---------- ---------- ---------- ------KVQV NVSPK-N--- ------IKTY STSTAYRGSL  51

Hs_TCN1_NP_001053.2            359 LSVMEKAQKM NDTIFGFTME -ERSW-GPYI TCIQGLCANN NDRTYWELLS GG-----EPL SQGAGSYVVR NGENLEVRWS  431
Hs_GIF_NP_005133.2             344 LVVLEEAQRK NPM-FKFETT M-TSW-GLVV SSINNIAENV NHKTYWQFLS GV-----TPL NEGVADYIPF NHEHITANFT  415
Hs_TCN2_NP_000346.2            356 EDVLKKAHEL G----GFTYE TQASLSGPYL TSVMGKAAG- -EREFWQLLR DPN----TPL LQGIADYRPK DGETIELRLV  425
Medaka_ENSORLT00000019254      55  LGAMRTLMDS DTN-FKFTYR EDPNY-GPHL ESINGLAGKD ADQTYWELLV MKPDGAITRP DVGIGCYIPS ANEKIIFNFT  132
Medaka_ENSORLT00000001408      82  IGAMKRLRKS NAN-FKFTYK EDLNY-GPYL ESINGVPGKT EDHTYWELLV IKPNGSVIIP DVGIGCYIPS PNEQILFNFT  159
Tilapia_ENSONIT00000013242     31  LGAMTRLRDS DAG-FTFTFS DNVNY-GPYL ESVNGVTGNN EAHTYWELLA NVTNGGFQRT EVGIGCIIPS PYQQIILNFT  108
Tetradon_ENSTNIT00000017613    58  LGGLRRLMKS NAG-FTFGYS EHPDY-GPFL ESVNGLAGSD RDRTYWELLV RTADGRLLRP DVGIGCYVPK PKDQIILNFT  135
Cavefish_ENSAMXT00000002854    43  LGAMRKIAAK TND-FKFTIR DDLNY-GPFL VSVNGVAGG- -DHTYWELLS KRANGTIIRP EVGVGCFIPD PDDTVILKYT  118
Stickleback_ENSGACT00000027033 59  LSALRRLQET QQD-FKFTVT VDPNF-GLFL ESVNGVAGSE SEQTYWEILS ESF-GEYTRL DVGIGCYQPV ADEHIILRFS  135
Platyfish_ENSXMAT00000005953   81  FGALTRLQET QPD-FKFTVT VDPNF-GLFL ESVNGVAGDE NQQTYWEILT ENS-GEYTRL DVGIGCYTPK ADEHIVLRFR  157
Amazonmolly_ENSPFOT00000010148 80  FGALTRLQET QPD-FKFTVK VDPNF-GLFL ESVNGVAGDE NEQTYWEILT ENS-GEYTRL DVGIGCYTPK ADEHIVLKFR  156
Tilapia_ENSONIT00000006763     59  LGALRRLHDA QHD-FKFTVK EDPNF-GLFL ESVNGVAGNK DEKTYWEILS ESS-GEFNRL DVGIGCYMPK ADEHIVLRYT  135
Tetraodon_ENSTNIT00000014896   58  MGALRRLQET QHN-FKFTVK WDPDF-GLFL ESVNGVAGNV HEQTYWEILS ESS-EEHRRI DLGLGCYKPK ANEHIILRFT  134
Codfish_ENSGMOT00000005936     30  LGALKRLQQT DPS-FRFTLK EDPDH-GLFL ESVNGVAGSG QAQSYWELLS ASAPGDPARL DAGIGCYKPK AGEHIILRLS  107
Cavefish_ENSAMXT00000003614    61  YGALTRLQDS SNG-FKFTVK IDPNL-GLYL ESVNGVAGSE AKHTYWQILS EHD-GTVTKL DVGVGCYQPK KDEHIILKYT  137
Zebrafish_ENSDART00000098273   60  FGVLNQLMES NAD-FKFSYT IHHTF-GIYL ESVNGLAGSD EDQTYWELLS EKS-GVVTRL EVGIGCYQVQ RDENLILRFT  136
Zebrafish_ENSDART00000147769   65  FGVLNQLQDS N-Q-LNFTYS ISKSY-GIFL ESVNGLAGST ENKTYWELLS KRE-RKTTRL NVGIGCYQPE RNENFIMNFT  140
Spottedgar_ENSLOCT00000004606  58  LGAMWRIQQA NSN-FSFETR DDINY-GPYL VSVNGVAGND TAHTYWQLLR YPK----TPL DRGVGCYIPK PNEHIILNFT  131
Spottedgar_ENSLOCT00000004639  62  LGAMLRIQQE NSN-FRFETR DDINY-GPYL VSVNGVAGND TAHTYWQLLR YPN----MPL DRGVGCYIPG ENEHIILRFT  135
Tilapia_ENSONIT00000020351     52  FGGLTRLKYS NQG-FNFQYI PNDDY-GPFL QSVNGLAGN- -SSYYWQLLS GK-----TPL DVGMGCYLPT ANEVVTLKYT  122

Hs_TCN1_NP_001053.2            432 KY------ 433
Hs_GIF_NP_005133.2             416 QY------ 417
Hs_TCN2_NP_000346.2            426 SW------ 427
Medaka_ENSORLT00000019254      133 KW------ 134
Medaka_ENSORLT00000001408      160 KW------ 161
Tilapia_ENSONIT00000013242     109 VW------ 110
Tetradon_ENSTNIT00000017613    136 RW------ 137
Cavefish_ENSAMXT00000002854    119 TW------ 120
Stickleback_ENSGACT00000027033 136 TWRRQ--- 140
Platyfish_ENSXMAT00000005953   158 TWNSTTVE 165
Amazonmolly_ENSPFOT00000010148 157 TWKSTTEE 164
Tilapia_ENSONIT00000006763     136 TWSPQQ-- 141
Tetraodon_ENSTNIT00000014896   135 KLQPR--- 139
Codfish_ENSGMOT00000005936     108 TWSKD--- 112
Cavefish_ENSAMXT00000003614    138 TWTKE--- 142
Zebrafish_ENSDART00000098273   137 TWATKK-- 142
Zebrafish_ENSDART00000147769   141 TWA----- 143
Spottedgar_ENSLOCT00000004606  132 TWDNLKRH 139
Spottedgar_ENSLOCT00000004639  136 TWD----- 138
Tilapia_ENSONIT00000020351     123 KI------ 124
```

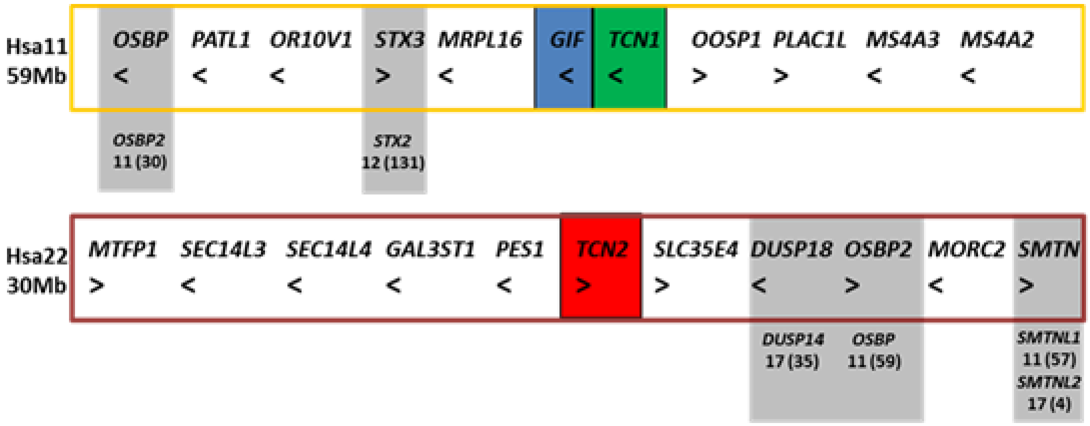
**Supplementary Fig.1. A** - Sequence alignment of holostean and teleost Tcn-like sequences with the 3 human cobalamin binders.



**Supplementary Fig. 1. B** - **S**ynteny maps of the Tcn-like *locus* in fish and the orthologous *locus* in human, Hsa – *H. sapiens*; Dre – *D. rerio*; Gac – *G. aculeatus*; Ola – *O. latipes* and Loc- *L. oculatus.*


96
Platyfish ENSXMAT00000005953
Amazon molly ENSPFOT00000010148
Tetraodon ENSTNIT00000014896
Stickleback ENSGACT00000027033
Tilapia ENSONIT00000006763
64
Codfish ENSGMOT00000005936
67
Cavefish ENSAMXT00000003614
Zebrafish ENSDART00000098273
93
Zebrafish ENSDART00000147769
Tilapia ENSONIT00000020351
53
77
Medaka ENSORLT00000019254
Medaka ENSORLT00000001408
100
Tilapia ENSONIT00000013242
Tetradon ENSTNIT00000017613
Cavefish ENSAMXT00000002854
Spottedgar ENSLOCT00000004639
Spottedgar ENSLOCT00000004606
0,1

**Supplementary Fig. 1. C** - Maximum Likelihood tree of Tcn-like genes, with 100 bootstrap replicates. Bootstrap values below 50 were removed.



Linkage group 17

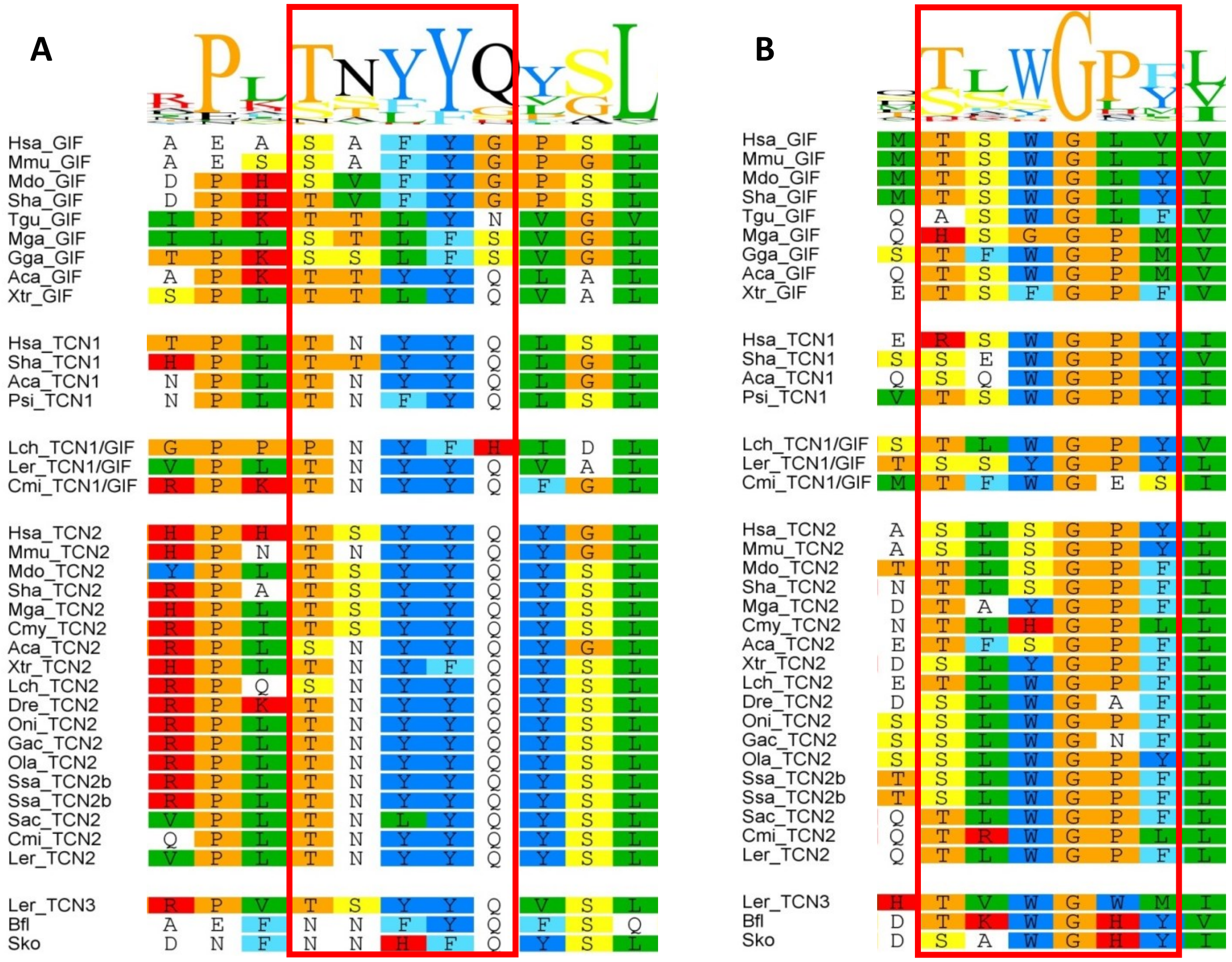
**17.1/3/5/8** 7.3

**22.1** 12.10

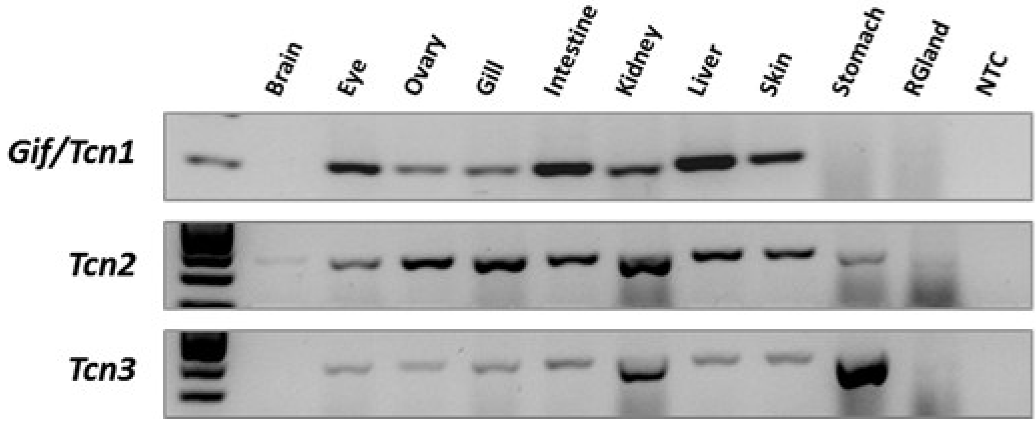
16.2

**11.5**

**Supplementary Fig. 2.** Paralogy analysis of the *Tcn1*, *Gif* and *Tcn2* human *loci* below (grey) corresponding paralogues of neighboring genes all mapping to the ancestral LG17 (see Putnam et al., 2008 for details of chromosome coordinates of linkage group 17).


A
Hsa_GIF A E A S A F Y G P S L
Mmu_GIF A E S S A F Y G P G L
Mdo_GIF D P H S V F Y G P S L
Sha_GIF D P H T V F Y G P S L
Tgu_GIF I P K T T L Y N V G V
Mga_GIF I L L S T L F S V G L
Gga_GIF T P K S S L F S V G L
Aca_GIF A P K T T Y Y Q L A L
Xtr_GIF S P L T T L Y Q V A L
Hsa_TCN1 T P L T N Y Y Q L S L
Sha_TCN1 H P L T T Y Y Q L G L
Aca_TCN1 N P L T N Y Y Q L G L
Psi_TCN1 N P L T N F Y Q L S L
Lch_TCN1/GIF G P P P N Y F H I D L
Ler_TCN1/GIF V P L T N Y Y Q V A L
Cmi_TCN1/GIF R P K T N Y Y Q F G L
Hsa_TCN2 H P H T S Y Y Q Y G L
Mmu_TCN2 H P N T N Y Y Q Y G L
Mdo_TCN2 Y P L T S Y Y Q Y S L
Sha_TCN2 R P A T S Y Y Q Y S L
Mga_TCN2 H P L T S Y Y Q Y S L
Cmy_TCN2 R P I T S Y Y Q Y S L
Aca_TCN2 R P L S N Y Y Q Y G L
Xtr_TCN2 H P L T N Y F Q Y S L
Lch_TCN2 R P Q S N Y Y Q Y S L
Dre_TCN2 R P K T N Y Y Q Y S L
Oni_TCN2 R P L T N Y Y Q Y S L
Gac_TCN2 R P L T N Y Y Q Y S L
Ola_TCN2 R P L T N Y Y Q Y S L
Ssa_TCN2b R P L T N Y Y Q Y S L
Ssa_TCN2b R P L T N Y Y Q Y S L
Sac_TCN2 V P L T N L Y Q Y S L
Cmi_TCN2 Q P L T N Y Y Q Y S L
Ler_TCN2 V P L T N Y Y Q Y S L
Ler_TCN3 R P V T S Y Y Q V S L
Bfl A E F N N F Y Q F S Q
Sko D N F N N H F Q Y S L
B
Hsa_GIF M T S W G L V V
Mmu_GIF M T S W G L I V
Mdo_GIF M T S W G L Y V
Sha_GIF M T S W G L Y I
Tgu_GIF Q A S W G L F V
Mga_GIF Q H S G G P M V
Gga_GIF S T F W G P M V
Aca_GIF Q T S W G P M V
Xtr_GIF E T S F G P F V
Hsa_TCN1 E R S W G P Y I
Sha_TCN1 S S E W G P Y V
Aca_TCN1 Q S Q W G P Y I
Psi_TCN1 V T S W G P Y I
Lch_TCN1/GIF S T L W G P Y V
Ler_TCN1/GIF T S S Y G P Y L
Cmi_TCN1/GIF M T F W G E S I
Hsa_TCN2 A S L S G P Y L
Mmu_TCN2 A S L S G P Y L
Mdo_TCN2 T T L S G P F L
Sha_TCN2 N T L S G P F I
Mga_TCN2 D T A Y G P F L
Cmy_TCN2 N T L H G P L L
Aca_TCN2 E T F S G P F L
Xtr_TCN2 D S L Y G P F L
Lch_TCN2 E T L W G P F L
Dre_TCN2 D S L W G A F L
Oni_TCN2 S S L W G P F L
Gac_TCN2 S S L W G N F L
Ola_TCN2 S S L W G P Y L
Ssa_TCN2b T S L W G P F L
Ssa_TCN2b T S L W G P F L
Sac_TCN2 Q T L W G P F L
Cmi_TCN2 Q T R W G P L L
Ler_TCN2 Q T L W G P F L
Ler_TCN3 H T V W G W M I
Bfl D T K W G H Y V
Sko D S A W G H Y I

**Supplementary Fig. 3.** Motif sequence alignment of *Tcn1*, *Gif* and *Tcn2*. (A) The TNNYQ motif in the α-domain suggested to form 4 hydrogen bonds with the corrinoid moiety in human *Tcn1* contributing for a high affinity towards Cbl;(B) the bulky hydrophobic residue region suggested to compensate for the missing nucleotide in corrinoids contributing for a lower specificity in binding.



**Supplementary Fig. 4.** *Tcn1/Gif*, *Tcn2* and *Tcn3* gene expression analysis in a tissue panel of *L. erinacea*.

# CHAPTER VI

## PROTEIN DIGESTION AND GASTRIC PROTEASES

# CHAPTER VI - PROTEIN DIGESTION AND GASTRIC PROTEASES

## V.1 THE EVOLUTION OF PEPSINOGEN C GENES IN VERTEBRATES: DUPLICATION LOSS AND FUNCTIONAL DIVERSIFICATION

L. FILIPE C. CASTRO*, MÓNICA LOPES-MARQUES*, ODETE GONÇALVES, JONATHAN M. WILSON

(*JOINT FIRST AUTHORS)

**PLOS ONE**

(DOI: 10.1371/journal.pone.0032852)

OPEN ACCESS Freely available online PLoS one

# The Evolution of Pepsinogen C Genes in Vertebrates: Duplication, Loss and Functional Diversification

**Luís Filipe Costa Castro[1]*[9], Monica Lopes-Marques[1][9], Odete Gonçalves[1,2], Jonathan Mark Wilson[1]***

**1** CIMAR Associate Laboratory, CIIMAR–Interdisciplinary Centre of Marine and Environmental Research, UPorto–University of Porto, Porto, Portugal, **2** Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal

## Abstract

***Background:*** Aspartic proteases comprise a large group of enzymes involved in peptide proteolysis. This collection includes prominent enzymes globally categorized as pepsins, which are derived from pepsinogen precursors. Pepsins are involved in gastric digestion, a hallmark of vertebrate physiology. An important member among the pepsinogens is pepsinogen C (*Pgc*). A particular aspect of *Pgc* is its apparent single copy status, which contrasts with the numerous gene copies found for example in pepsinogen A (*Pga*). Although gene sequences with similarity to *Pgc* have been described in some vertebrate groups, no exhaustive evolutionary framework has been considered so far.

***Methodology/Principal Findings:*** By combining phylogenetics and genomic analysis, we find an unexpected *Pgc* diversity in the vertebrate sub-phylum. We were able to reconstruct gene duplication timings relative to the divergence of major vertebrate clades. Before tetrapod divergence, a single *Pgc* gene tandemly expanded to produce two gene lineages (*Pgbc* and *Pgc2*). These have been differentially retained in various classes. Accordingly, we find *Pgc2* in sauropsids, amphibians and marsupials, but not in eutherian mammals. *Pgbc* was retained in amphibians, but duplicated in the ancestor of amniotes giving rise to *Pgb* and *Pgc1*. The latter was retained in mammals and probably in reptiles and marsupials but not in birds. *Pgb* was kept in all of the amniote clade with independent episodes of loss in some mammalian species. Lineage specific expansions of *Pgc2* and *Pgbc* have also occurred in marsupials and amphibians respectively. We find that teleost and tetrapod *Pgc* genes reside in distinct genomic regions hinting at a possible translocation.

***Conclusions:*** We conclude that the repertoire of *Pgc* genes is larger than previously reported, and that tandem duplications have modelled the history of *Pgc* genes. We hypothesize that gene expansion lead to functional divergence in tetrapods, coincident with the invasion of terrestrial habitats.

**Citation:** Castro LFC, Lopes-Marques M, Gonçalves O, Wilson JM (2012) The Evolution of Pepsinogen C Genes in Vertebrates: Duplication, Loss and Functional Diversification. PLoS ONE 7(3): e32852. doi:10.1371/journal.pone.0032852

**Editor:** David Liberles, University of Wyoming, United States of America

**Received** November 8, 2011; **Accepted** February 3, 2012; **Published** March 9, 2012

**Funding:** These authors have no support or funding to report.

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

* E-mail: filipe.castro@ciimar.up.pt (LFC); wilson.jm.cimar@gmail.com (JMW)

[9] These authors contributed equally to this work.

## Introduction

Pepsinogens, the precursors of pepsins, are a group of aspartic proteases involved in the specific hydrolysis of peptides. Typically, they show a high and localized expression in the stomach due to their crucial role in protein digestion. After secretion into the gastric lumen, the pepsinogens are activated into pepsins by the action of hydrochloric acid, which alters their structural conformation [1]. The activation involves the autocatalytic cleavage of the prosegment from the N-terminus of the enzyme [1].

A wide diversity of pepsinogen genes is found in mammalian species. The five group nomenclature distinguishes pepsinogen A (*Pga*), B (*Pgb*), C (*Pgc*), F (*Pgf*), and prochymosin (*Cym*) [2,3]. The various pepsinogen gene families are thought to have emerged from a common intracellular aspartic protease through gene duplication, though the exact duplication timings and processes are presently unknown [4,5]. Within each pepsinogen family, gene numbers vary significantly between species and gene family. For example, *Pga* has three gene copies in humans, while a single copy is found in the opossum [6]. An extreme case of lineage-specific gene expansion was recently determined in the orangutan where fourteen different *Pga* cDNAs were found, corresponding to a minimum of eight *loci* [7]. This is in sharp contrast to the condition observed in the *Pgc* gene family, which is mostly considered single copy [3]. Therefore, *Pgc* has been suggested as a reliable molecular marker in species phylogenetic analysis [2,8]. This distinctive feature of *Pgc* is apparently corroborated by the characterization of single cDNAs in vertebrate classes such as teleosts [9] amphibians [10] and birds [11]. Contradictorily, Ordoñez et al. [6] have suggested the presence of extra *Pgc-like* sequences in some vertebrate species. Nevertheless, no phylogenetics or comprehensive species sampling was performed thus preventing elaboration on duplication timings/processes or evolutionary history. For example, it was argued that *Pgb* and *Pgc* derived from tandem gene duplication in the therian mammalian ancestor [6], a proposal impossible to confirm without phylogeny. In this study, we set out to investigate the Pgc, a gene family which together with other pepsinogen isoforms is fundamental for the vertebrate gastric function. We take an approach that combines phylogenetics and comparative genomics to unravel a complex evolutionary pathway of *Pgc* in the vertebrate sub-phylum. We find that contrary to previous reports, the diversity of the *Pgc* gene family is broader with various episodes of gene duplication and loss, particularly in tetrapods. Based on the current findings we recommend a new gene nomenclature for *Pgc* genes which incorporates gene duplication history and phylogenetic distribution.

## Methods

### Identification of *Pgc* genes

*Pgc* sequences were identified in the Ensembl and GenBank databases for the following species with genome sequences available: *Homo sapiens* (human), *Pan troglodytes* (common chimpanzee), *Gorilla gorilla* (Gorilla), *Loxodonta fricana* (African savanna elephant), *Sus scrofa* (pig), *Mus musculus* (mouse), *Rattus norvegicus* (brown rat), *Monodelphis domestica* (opossum), *Xenopus tropicalis* (western clawed frog), *Anolis carolinensis* (anolis), *Gallus gallus* (chicken), *Meleagris gallopavo* (turkey), *Tetraodon nigroviridis* (green spotted puffer), *Takifugu rubripes* (pufferfish), *Danio rerio* (zebrafish), *Oryzias latipes* (medaka) and *Gasterosteus aculeatus* (stickleback). To identify non-annotated genes Blastp searches were performed using the human PGC protein sequence. Blast searches to EST databases (when available) were also implemented. Sequences previously described in organisms (teleosts) without genome sequences were also incorporated in the phylogenetic analysis. Accession numbers for the sequences are listed in Table 1. The alignment provided in Fig. 1 was performed in Geneious V5.4.6 [12] with the Clustal plugin (settings below).

### Phylogenetic analysis

PGC amino acid sequences were aligned in ClustalX 2.0.11 with standard settings (Gonnet weight matrix, gap opening = 10 and gap extension = 0.2) [13,14]. All positions containing gaps and missing data were eliminated. The final dataset involved 42 amino acid sequences and 339 positions. The evolutionary history was inferred using two methods. A Neighbor-Joining (NJ) tree [15] was reconstructed using standard settings with ClustalX 2.0.11 [13,14]. The robustness of the tree was assessed through 1000 bootstrap replicates of the data. The same alignment was also used to generate a Maximum likelihood tree (ML). The evolutionary model was derived from ProtTest (LG+I+G+F) [16]. The ML tree was reconstructed using PhyML online [17] with the amino acid frequency (equilibrium frequency), proportion of invariable sites and gamma-shape (4 rate substitution categories) for the amino acid substitution rate heterogeneity parameters estimated from the dataset. Bootstrap analysis (1000 replicates) was carried out to determine the robustness of the tree. The pufferfish *Pga* sequence was used as an outgroup in TreeView 1.6.6.

### Comparative genomics and neighboring gene families

The chromosomal location of the *Pgc* genes and the flanking gene families was collected from the Ensembl and GenBank databases. The human *PGC* and *PGBψloci* were used as the tetrapod genomic models for comparison with the stickleback. Information on the evolutionary history (orthologous *versus* paralogous) of the gene families flanking human and stickleback *Pgc loci* was collected from the Ensembl paralogue and orthologue prediction pipeline.

### Structural comparative modeling

The crystal structure of the *H. sapiens* progastricsin (1HTR) [18] was used as a template for 3D modeling. To predict structure a Modeller algorithm [19] available at HHpred was used [20,21] The predicted structural models were evaluated using modeller output Verify 3D [22] and in all cases accurate structures were achieved. Structural visualization and analysis was performed using Open-Source PyMOL V1.3. (academic version) [23].

## Results

### An unexpected diversity of *Pgc* genes in tetrapods

We initiated this research by first establishing whether the single copy status of *Pgc* is a typical feature of this gene family. All *Pgc* sequences retrieved from database search are listed in Table 1. In the investigated mammalian species we find three dissimilar gene complements. Human, gorilla, chimpanzee, mouse and rat all have a single *Pgc* sequence, while the African elephant and pig have two. In humans a second *Pgc-like* sequence, named *Pgb*, is also found though this is a pseudogene [6]. In contrast, the opossum has five identifiable open reading frames (ORFs) with similarity to the *Pgc* gene family. Birds represented here by the chicken and turkey, have two *Pgc-like* sequences, whereas in the lizard we have uncovered three. A fourth sequence is present in the reptile genome but the current assembly indicates a frameshift mutation in the eighth exon producing a truncated protein (Figure S1). Whether this is the first step of pseudogenization or a sequencing error remains to be determined. The western clawed frog has the same number of *Pgc-like* sequences to that found in the opossum. In teleosts we find two distinct situations. Whilst in the stickleback a single representative of *Pgc* was recovered, other fish species with available genome sequences have no identifiable hits to *Pgc*.

We next performed sequence alignment of the all the collected sequences (Fig. 1). In all sequences three distinct regions can be distinguished: the signal peptide (1–16 human; PGC coordinates used hereafter, Fig. 1 red bar); the activation peptide (residues 16–59, Fig. 1 blue bar), and the active enzyme moiety (residues 59–388). In the activation segment (prosegment) highly conserved residues were observed (Fig. 1 yellow boxes 1). These residues (pLeu7, pSer12, pArg14, pGly21 and pLys37 - p prosegment numbering) are also conserved in PGA and PGY suggesting that they play an important role in the activation segment [3,24]. In fact pepsinogens are activated by the cleavage of the prosegment. There are two major cleavage sites in human PGC, one located between pPhe26 and pLeu27 and the second located at the last residue of the prosegment pLeu43 and the first residue of the enzyme moiety Ser1 [1,3,25]. In a neutral pH the prosegment is coupled to the enzyme moiety by electrostatic interactions and hydrogen-bonds, pLys37, pTyr38 and Tyr9 (Fig. 1 green boxes) bind to the catalytic aspartates (Fig. 1 black boxes "+") [1,26,27,28] In an acidic pH environment acidic residues in the enzyme moiety become protonated disrupting electrostatic interactions with the prosegment (which has a basic character), releasing the prosegment for proteolytic cleavage and enzyme activation [3,29,30]. In fish pepsinogens a deletion of several residues in the prosegment is observed (Fig. 1 activation segment, lower black bar) leads to a decrease in the number of basic residues in the prosegment, and given the PI values for each enzyme region (Table 1), we deduced that the activation of fish pepsinogens occurs in conditions that are comparatively more alkaline. In accordance, the analysis of the PI of each enzyme moiety supports this observation (Table 1). While tetrapod *Pgc* have values below 3,5 (with the exception of *Pgb*), teleost *Pgc* are mostly above 3,6 and closer to 4, suggesting distinct activation conditions. Furthermore, the cleavage of the prosegment can be completed in the sequential pathway or in a direct pathway [1]. In human the sequential pathway involves an initial cleavage between pPhe26-pLeu27

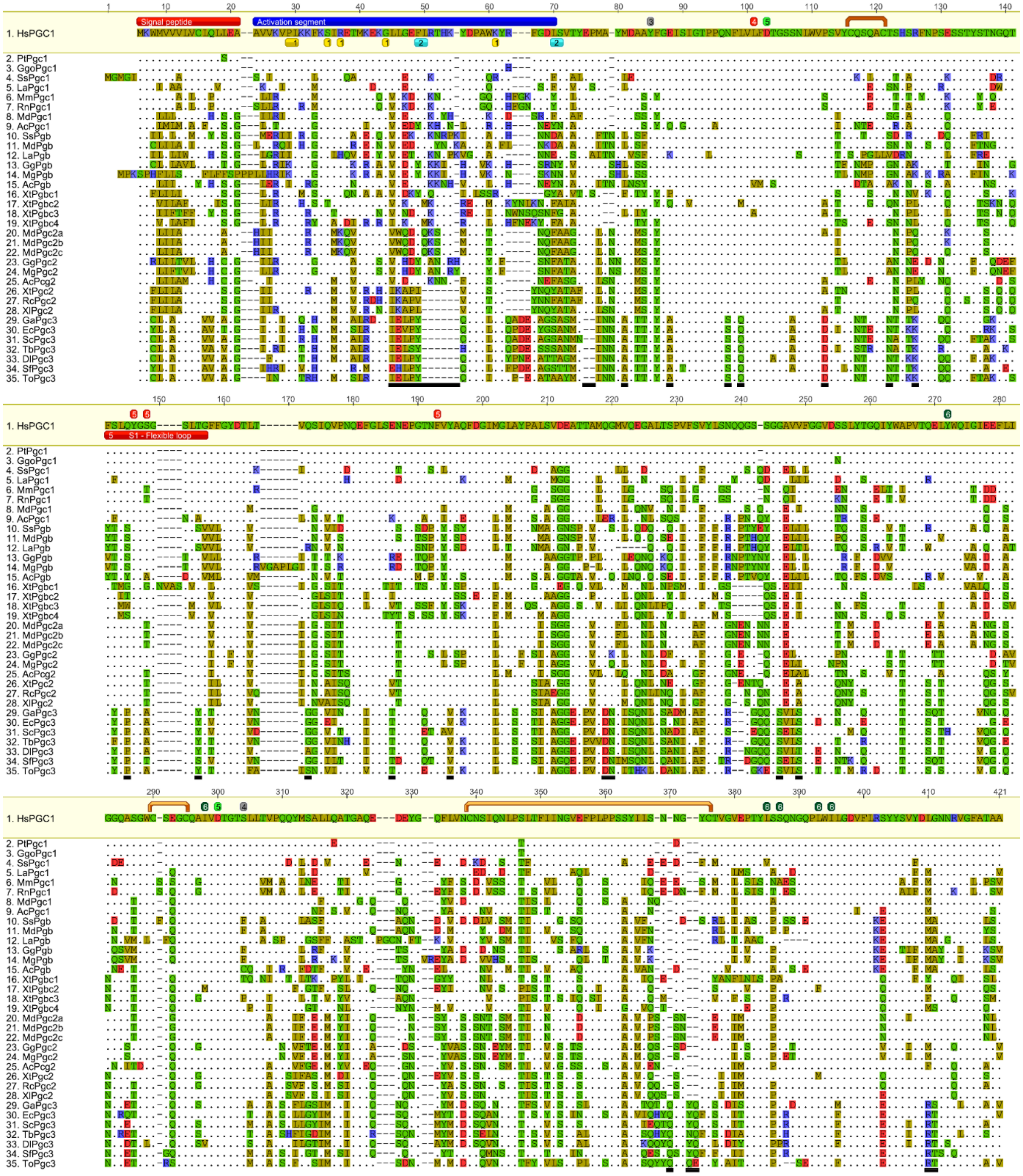

Signal peptide
Activation segment
S1 - Flexible loop
1. HsPGC1
2. PtPgc1
3. GgoPgc1
4. SsPgc1
5. LaPgc1
6. MmPgc1
7. RnPgc1
8. MdPgc1
9. AcPgc1
10. SsPgb
11. MdPgb
12. LaPgb
13. GgPgb
14. MgPgb
15. AcPgb
16. XtPgbc1
17. XtPgbc2
18. XtPgbc3
19. XtPgbc4
20. MdPgc2a
21. MdPgc2b
22. MdPgc2c
23. GgPgc2
24. MgPgc2
25. AcPgc2
26. XtPgc2
27. RcPgc2
28. XlPgc2
29. GaPgc3
30. EcPgc3
31. ScPgc3
32. TbPgc3
33. DlPgc3
34. SfPgc3
35. ToPgc3
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400
410
421

**Figure 1. Multiple sequence alignment of vertebrate *Pgc-like* sequences performed in Geneious V5.5.6 using Clustal plugin with Gonnet scoring matrix and the following parameters: Gap opening = 10, Gap extension = 0.2.** The red bar indicates signal peptide, blue bar activation segment or propeptide, yellow boxes highly conserved residues of the propeptide [3], light green boxes residues (pLys37 pTyr38 and Tyr9) involved in interactions that block access to the catalytic aspartates at neutral pH [1] black boxes "+" conserved catalytic aspartates (Asp32 and Asp217), orange bridges six conserved cysteines involved in the formation of disulphide linkages (Cys45, Cys50, Cys208, Cys212, Cys251, Cys284), grey boxes residues reported to be involved in pepsinogen B substrate specifity [13] and underlining black boxes sequence features specific to fish pepsinogens (All coordinates are relative to human PGC).
doi:10.1371/journal.pone.0032852.g001

bond followed by the cleavage of the prosegment from the active enzyme between residues pLeu43 and Ser1 (Fig. 1 Green boxes). Given that pLeu27 is deleted in fish and western clawed frog prosegment and pPhe26 is not conserved in fish, lizard, birds and western clawed frog, a direct or distinct activation pathway is expected for these species [1,3,25].

All PGC-like sequences analysed here present the highly conserved catalytic-site aspartates, Asp32 and Asp217, character-

**Table 1.** List of accession numbers for all the *Pgc* sequences used in the phylogenetic analysis.

| Species | Accession number | Name | Pro-segment PI | Moietyc PI |
|---|---|---|---|---|
| *Homo sapiens* | NP_002621 | *PGC1* | 10.67 | 3.35 |
| *Pan troglodytes* | XP_518465 | *Pgc1* | 10.67 | 3.31 |
| *Gorilla gorilla* | ENSGGOP00000002651 | *Pgc1* | 10.48 | 3.35 |
| *Loxodonta africana* | ENSLAFP00000025164 | *Pgb* | 10.11 | 4.16 |
|  | ENSLAFP00000014376 | *Pgc1* | 10.60 | 3.54 |
| *Sus scrofa* | XP_003355296 | *Pgb* | 10.28 | 3.62 |
|  | XP_003128442 | *Pgc1* | *10.67* | 3.51 |
| *Mus musculus* | NP_080249 | *Pgc1* | *10.47* | 3.50 |
| *Rattus norvegicus* | NP_579818 | *Pgc1* | *10.40* | 3.46 |
| *Monodelphis domestica* | NP_001028152 | *Pgb* | *10.28* | 3.68 |
|  | XP_001370482 | *Pgc2c* | *10.92* | 3.26 |
|  | XP_001370462 | *Pgc2b* | *10.92* | 3.38 |
|  | XP_001370404 | *Pgc1* | *10.14* | 3.49 |
|  | XP_001370435 | *Pgc2a* | *10.92* | 3.38 |
| *Meleagris gallopavo* | ENSMGAP00000018167 | *Pgb* | *10.91* | 4.83 |
|  | ENSMGAP00000006238 | *Pgc2* | *10.15* | 3.23 |
| *Gallus gallus* | XP_425832 | *Pgb* | *11.00* | 4.24 |
|  | NP_990208 | *Pgc2* | *10.29* | 3.14 |
| *Anolis carolinensis* | XP_003220378.1 | *Pgb* | *10.06* | 3.61 |
|  | Gene ID:100567329 | *ΨPgb* | n.a | n.a |
|  | XP_003220379.1 | *Pgc1* | *9.61* | 3.92 |
|  | XP_003220377.1 | *Pgc2* | *10.35* | 3.09 |
| *Xenopus tropicalis* | XM_002932980 | *Pgc2* | *10.54* | 3.21 |
|  | XM_002932982 | *Pgbc1* | *10.64* | 3.68 |
|  | NM_001030432 | *Pgbc2* | *10.84* | 3.56 |
|  | NM_001015682 | *Pgbc3* | *11.08* | 3.52 |
|  | NM_001032309 | *Pgbc4* | *10.78* | 3.50 |
| *Xenopus laevis* | AB045379 | *Pgc2* | *10.54* | 3.26 |
| *Rana catesbeiana* | M73750 | *Pgc2* | *10.63* | 3.28 |
| *Gasterosteus aculeatus* | ENSGACG00000012388 | *Pgc3* | *9.06* | 3.85 |
| *Epinephelus coioides* | EU136029 | *Pgc3* | *8.99* | 4.17 |
| *Siniperca chuatsi* | FJ463157 | *Pgc3* | *9.77* | 3.74 |
| *Trematomus bernacchii* | AJ550952 | *Pgc3* | *9.08* | 4.53 |
| *Dicentrarchus labrax* | EF690286 | *Pgc3* | *9.63* | 4.08 |
| *Salvelinus fontinalis* | AF275939 | *Pgc3* | *9.09* | 3.69 |
| *Thunnus orientalis* | AB440202 | *Pgc3* | *9.86* | 4.08 |

Also shown is the isolelectric point (PI) of the pro-segment and of the pepsin moiety.
doi:10.1371/journal.pone.0032852.t001

istic of the aspartic protease family (Fig. 1, black boxes "+") and the six conserved cysteines reported to be involved in the formation of three bisulfide bridges (Fig. 1 orange bridges, Cys45–Cys50; Cys208–Cys212 and Cys251–Cys284). Although the elephant PGB presents a Serine at position 45 corresponding to the first Cysteine, this bridge (Cys45–Cys50) has been reported as unessential in the correct protein folding [31].

The catalytic-site aspartates are found in a substrate binding cleft in the enzyme moiety, and bordered by S1 and S1′ subsites (Fig. 1 red boxes "#" and dark green boxes "*", respectively). These subsites are involved in the binding of the substrate to the enzyme and have been reported to play an essential role in substrate specificity [32]. The S1 subsite is highly conserved in all PGC-like sequences and is located near the Asp32, presenting a flexible loop (Fig. 1, Light blue box "S1- Flexible loop") formed by several residues namely Phe71-Gly81, Leu30 Tyr75, Ser77 and Phe112 [1,3].

The S1′subsite, which is located in the neighbourhood of Asp217 and is formed by the following residues Tyr190, Ile215, Leu293, Ser295, Leu301 and Ile303, (Fig. 1 dark green boxes "*") and is also highly conserved between the analysed sequences, although neighbouring residues may vary. The S1 and S1′ subsites show distinct degrees of conservation. In fact the S1 subsite is comparatively less conserved. This may be related to the fact that the S1 subsite has been reported to play an important role in substrate binding [32]. Therefore, a higher residue variation at this site suggests diversification in the substrate cleft in order to accommodate distinct substrates.

Considering the S1′ subsite in detail we can observe residue patterns that are characteristic of a determined pepsinogen group. For example all PGB-like sequences present a valine residue at position 82 (Fig. 1 light blue box "S1 Flexible loop") along with this amino acid, PGB-like sequences tend to present a threonine at position 72 and a serine at position 74. Concerning the fish PGC (PGC), it is possible to observe that in the S1 subsite there are several distinctive features. All fish PGC sequences present an aromatic residue tyrosine at position 72 (phenylalanine in the case of Tb), which in other species is generally a serine or a threonine. In addition, at position 81 fish PGC have a tyrosine or serine in contrast to a highly conserved threonine in all other sequences and an exclusive proline is found at position 74. This subtle distinct residue composition in the S1 subsite of the PGC-like sequences may lead to distinct network of hydrogen bonds likely shifting substrate specificities.

## Phylogenetic analysis indicates multiple events of *Pgc* gene duplication

The finding of numerous *Pgc-like* sequences per species is surprising given previous reports arguing its single copy condition [3]. To clarify the evolutionary relationships between the various sequences as well as the duplication timings, we next constructed phylogenetic trees with NJ and ML (Fig. 2A and Fig. 2B). Both tree reconstructions show similar relationships between the retrieved sequences, with some differences. We find that classical *Pgc* (hereafter renamed *Pgc1*) and *Pgb* genes are found not only in the mammalian lineage as previously suggested [6]. A strongly supported *Pgb* clade includes a sequence also from birds and the reptile with both phylogenetic methods. The phylogenetic placement of one anolis sequence gave contradictory results. In the NJ tree the sequence is basal to the mammalian eutherian *Pgc1* (bootstrap 623), while in the ML tree the same sequence it comes basal to the *Pgb* clade (619 bootstrap). Based on the ML tree this sequence could represent a new gene lineage, which emerged in the ancestor of amniotes but was lost subsequently in birds, and mammals, with the reciprocal loss of *Pgc1* in reptiles. Four genes from the western clawed frog form an independent group which is basal to *Pgb* and *Pgc1* clades in both analyses, thus indicating that the *Pgb/Pgc1* duplication postdates amphibian divergence (Fig. 2). Both trees also display a third gene lineage found exclusively in birds (one gene), reptiles (one gene), amphibians (one gene) and marsupials (three genes), but not in eutherian mammals (Fig. 2A and Fig. 2B). The amphibian and bird case is particularly relevant, since these sequences were reported as *Pgc1* orthologues [10,11]. However, our analysis clearly indicates that these gene sequences belong to a distinct gene lineage. The fourth anolis *Pgc-like* sequence which has a frameshift mutation in the eight exon, robustly groups with the reptile *Pgb* sequence (Figure S1), indicative of lineage specific duplication followed by loss. Finally, teleosts outgroup the full tetrapod gene collection. Based on the phylogenetic analysis, we introduce here a new gene nomenclature for *Pgc* genes in tetrapods which takes into account the evolutionary relationships between the various genes (Table 1 and Fig. 2). Thus, we maintain the designation for *Pgb* but modify the previous *Pgc* to *Pgc1*. The basal amphibian clade we name *Pgbc* (with an *1* to *4* nomenclature to designate each independent gene), and the new gene lineage emerging from the phylogeny is designated *Pgc2*. Teleost *Pgc* genes are named *Pgc*. In summary, our search identified at least four evolutionary independent gene lineages in tetrapods, *Pgb*, *Pgcb*, *Pgc1* and *Pgc2*. Independent gene expansions are observed at specific lineages in the amphibian and marsupial clades. Taking into account the duplication patterns emerging from the phylogenetic analysis, we anticipate also that various independent events of gene loss have taken place. That is the case of *Pgb* in some mammalian species (e.g. human), *Pgc2* in eutherian mammals, and *Pgc1* in birds.

## Tetrapod *Pgc* genes reside in a gene cluster

We next examined the genomic location of *Pgc* genes in tetrapods (Fig. 3), since it can provide powerful insights with respect to gene origin and loss. We find that *Pgc loci* are extremely well conserved between the various species, with two distinct settings. In basal tetrapods such as amphibians, chicken and anolis, we find the full *Pgc* gene portfolio mapping to a single location; while in mammals, *Pgc1* and *Pgb* genes reside at two distinct genomic locations (Fig. 3). We conclude that the expansion of the *Pgc* gene lineage in the ancestor of tetrapods occurred through tandem gene duplications. We further find that the *Pgb* translocation to a separate genomic *locus* is a more recent event which took place in the ancestor of mammals, in contrast to previous suggestions [6], since *Pgb* maps to the same *locus* in both the opossum and pig (similar to the *Pgb* pseudogene in humans).

In contrast to tetrapods a single *Pgc* gene sequence has been described in various teleost species [9], a conclusion we now extent to stickleback. In this species, we find a single *Pgc* gene localizing to Group XIX (Fig. 4). A close inspection of the gene families flanking *GacPgc* shows no evidence of syntenic conservation in comparison to the tetrapod *Pgc locus*. We find for example that the stickleback orthologues of *Frs3* and *Tfeb* which outflank the *Pgc* gene cluster in tetrapods localize to Sca_27 in stickleback (not shown). Thus, *Pgc* has been apparently translocated from its original position in either tetrapods or teleosts and is of no evolutionary meaning. Mapping information from cartilaginous fish and pre-3R teleost species may provide insightful information on this issue. Except for the stickleback, we found no *Pgc-like* sequence in other teleost species with full genome sequences. To confirm the loss of *Pgc* sequences we analysed the composition of the *GacPgc locus* in zebrafish, medaka, pufferfish, and green pufferfish (Fig. 4). This approach confirms that neither of these

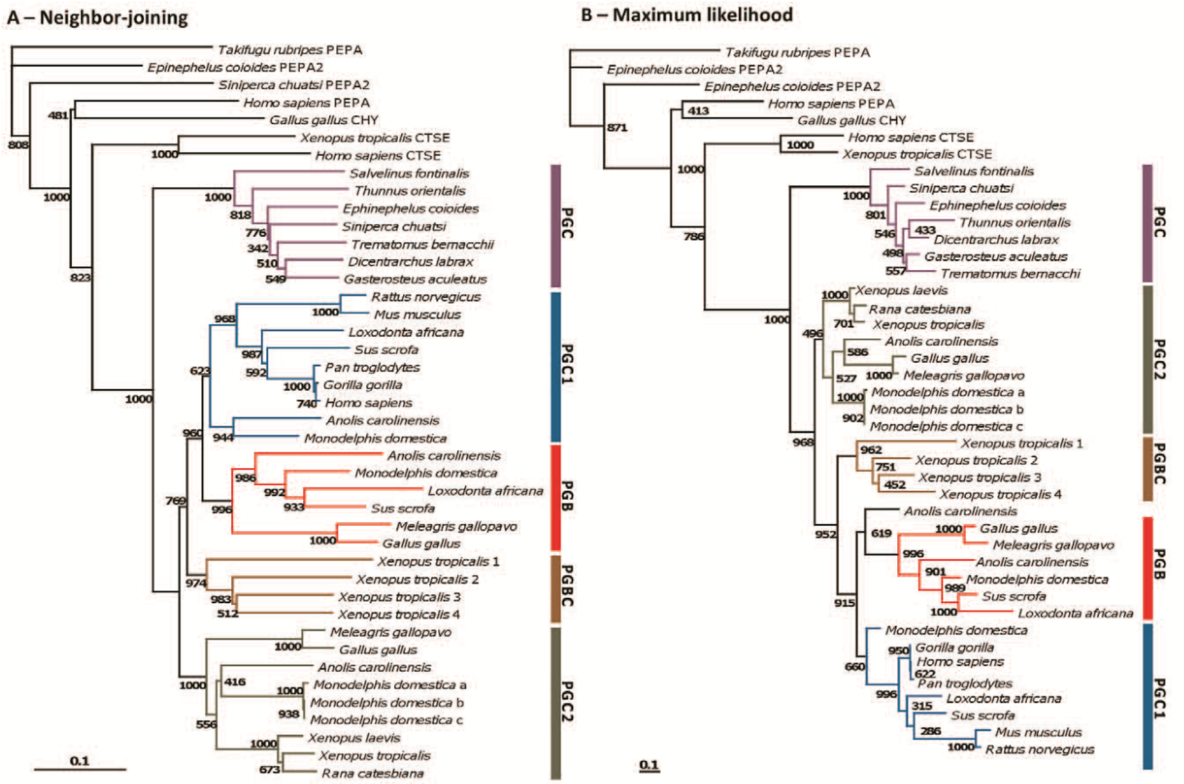


**Figure 2. Neighbor-Joining (A) and Maximum likelihood (B) tree of the *Pgc* gene family.** Values at nodes are bootstrap values (1000).
doi:10.1371/journal.pone.0032852.g002

species has a *Pgc* sequence in the genome (nor evidence for pseudogenization), despite the conservation of *locus* composition and organization.

## Discussion

Here we analyse the evolutionary history of a gene family involved in the vertebrate gastric function, the *Pgc*, to find that extensive gene duplication and loss occurred in vertebrate classes. Our research begun by inquiring a long held premise that *Pgc* is a single copy gene family in vertebrate species [3,9,10,11]. By taking an exhaustive search into various vertebrate genomes, we demonstrate that significant discrepancy in *Pgc* complements exists between species. For example, we find no *Pgc-like* gene in some teleost species (e.g. medaka), while up to five genes are found in the opossum and the western clawed frog. We next undertook a combination of phylogenetics and chromosomal gene location (and their neighbouring gene families) to reconstruct gene duplication timings and processes, relative to the divergence of major vertebrate classes. Our analysis supports an evolutionary scenario where tandem gene duplication and gene loss have dynamically taken place in the tetrapod lineage (Fig. 5). Consequently, we introduce a new gene nomenclature that incorporates the phylogenetic findings. Before the diversification of tetrapods, a gene duplication gave origin to two tandem paralogues, *Pgbc* and *Pgc2*. Preliminary data from the genome sequence of the coelacanth (*Latimeria chalumnae*) suggests that the duplication postdates the divergence of this basal Sarcopterygii lineage. *Pgc2* was maintained in most tetrapod species, but not in placental mammals. Episodes of lineage specific expansion were also observed in the opossum. As for the *Pgbc* gene, it expanded independently in the western clawed frog to held four gene copies (Fig. 5). Following the separation of amphibians but before amniote divergence, the *Pgbc* gene tandem duplicated to originate *Pgc1* and *Pgb* (Fig. 4), the latter being translocated from the *Pgc locus* in mammals. *Pgc1* was retained in most species, but not in the chicken and turkey, while *Pgb* experienced events of loss in some mammalian species, namely humans (Fig. 5). One anolis sequence (*Pgc1*) is inconsistently placed with both phylogenetic methods. In the NJ tree, it groups with the opossum *Pgc1* and basal to all other mammalian *Pgc1* genes. However, in the ML tree the same sequence is basal to the *Pgb* clade. If we consider the ML tree pattern correct, then this new gene represents a new lineage which emerged in the ancestor of amniotes but was lost subsequently in birds (1 event), and mammals (second event), plus the loss of *Pgc1* in the reptile. In contrast, the NJ tree requires less duplication and loss events. Thus, we consider more parsimonious to conclude that the anolis sequence is a true *Pgc1* gene. The position of the sequence in the *Pgc* gene cluster is also in agreement with this interpretation. Although this is only indicative evidence, this gene maps on the side of *Tfeb*, just as is observed in other species.

Surprisingly, *Pgc* has not been retained in every examined fish species (Fig. 4, Fig. 5). We find that some teleosts have no representative of *Pgc* in their genome, though other pepsinogen families can be found [33]. Documenting patterns of gene loss is of extreme relevance, particularly for the understanding of phenotypic evolution [34]. Furthermore, gene loss has been correlated with the evolution of functional changes in surviving gene family

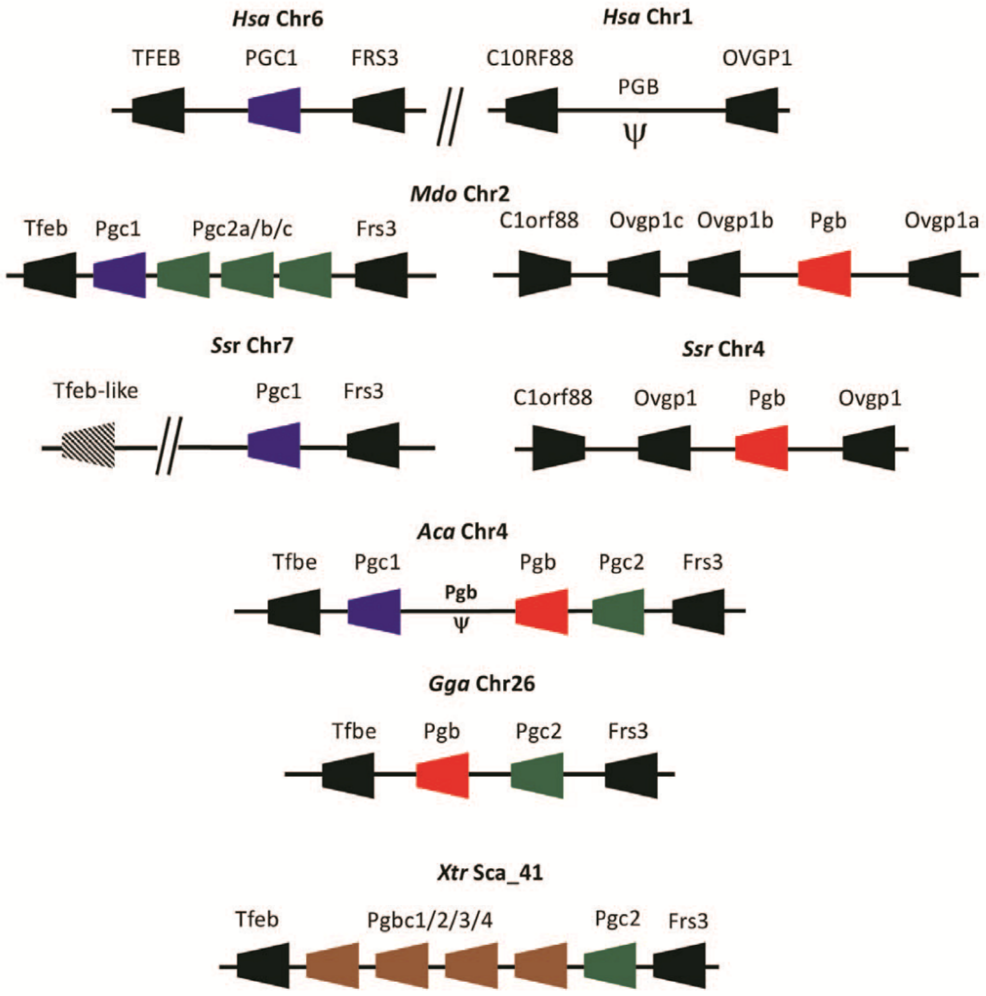


**Figure 3. Synteny map of *Pgc loci* in tetrapods.** Dashed gene box represents a TFEB partial ORF. Arrow head indicates gene orientation and *ψ* indicates pseudogene.
doi:10.1371/journal.pone.0032852.g003

members [35]. Currently, it is unclear whether the loss of *Pgc* genes in some teleosts and that of other *Pgc-like* lineages in tetrapods (e.g. *Pgb*) affected the evolution of additional pepsinogen gene family members (e.g. *Pga*), as well as, the gastric function.

Gene duplication is major source of morphological and functional innovation. The retention of the descendent gene copies can lead to the partitioning of ancestral functions or alternatively to the emergence of novel roles [36,37]. The finding of different *Pgc* gene lineages (and complements) in vertebrate classes suggests that functional divergence took place between isoforms. It has been argued that pepsinogen gene expansion, namely in *Pga*, enabled the appearance of proteins with different specificities, being advantageous for effective gastric digestion [7]. Experimental assays with PGC1 and PGB have hinted at distinct functional profiles. For example, porcine PGB hydrolytic activity towards haemoglobin is residual, when compared to human PGC1 [3]. Also, teleost *Pgc* shows minor specific activity towards haemoglobin as well [9]. The analysis of the dog PGB peptide cleavage capacity demonstrated preference for Phe–X bonds while PGC1 cleaves Tyr-X bonds [3,24]. In PGB, Tyr13 and Phe221 were shown to be crucially involved in substrate specificity; molecular modeling of pepsin B demonstrated that these residues lead to distinct network of hydrogen bonds and consequently accommodate different substrates in the binding cleft [25]. For other *Pgc* isoforms described here no experimental data is yet available. However, despite the high conservation degree at proposed critical enzymatic residues some sequence differences are discernible (Fig. 1). Several non-conserved residues located in the S1 and S1′subsites and their vicinities suggest subtle structural changes, namely in the enzyme structure, exposure of the active aspartic residues and in the general architecture of the binding cleft (Figure S2 and details therein). Thus, we propose the *Pgc* gene expansion was accompanied by the acquisition of novel substrate specificities.

The expansion of *Pgc* gene family finds parallel in other pepsinogen gene families, namely *Pga* [3,7]. In hominoids, two separate *Pga* lineages (and isoform numbers) have been described, *Pga1* and *Pga2* [7]. Interestingly, *Pga1* and *Pga2* have rather distinct PI values suggesting activation at different pHs, analogous to our findings in *Pgc* (Table 1). Also, it has been argued that the expansion of *Pga* might be advantageous for gastric digestion [7],

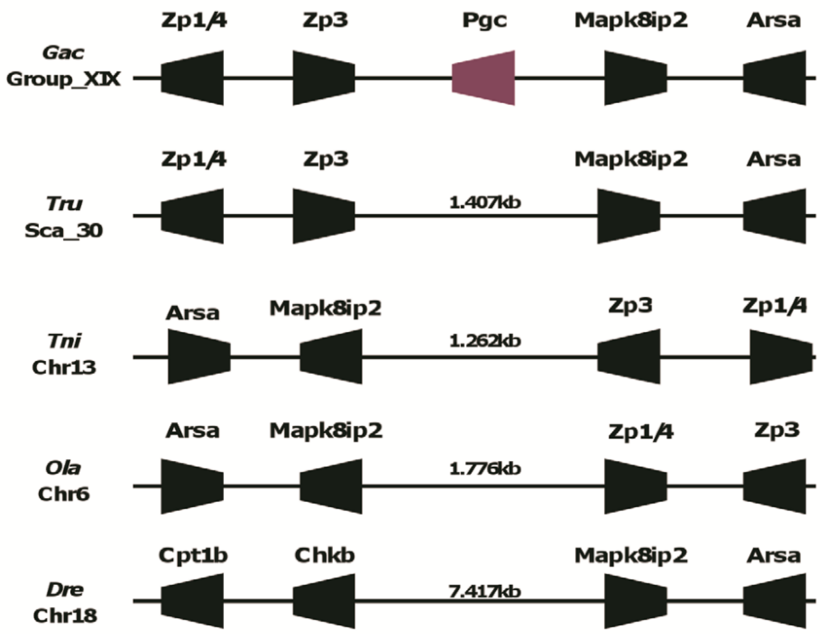


**Figure 4. *Pgc loci* are conserved in teleost species, and indicate gene loss in some lineages.** Gac – *G. aculeatus*, Tru- *T. rubripes*, Tni- *T. nigroviridis*, Ola- *O. latipes* and Dre- *D. rerio*. Numbers are distances between genes in Kb.
doi:10.1371/journal.pone.0032852.g004

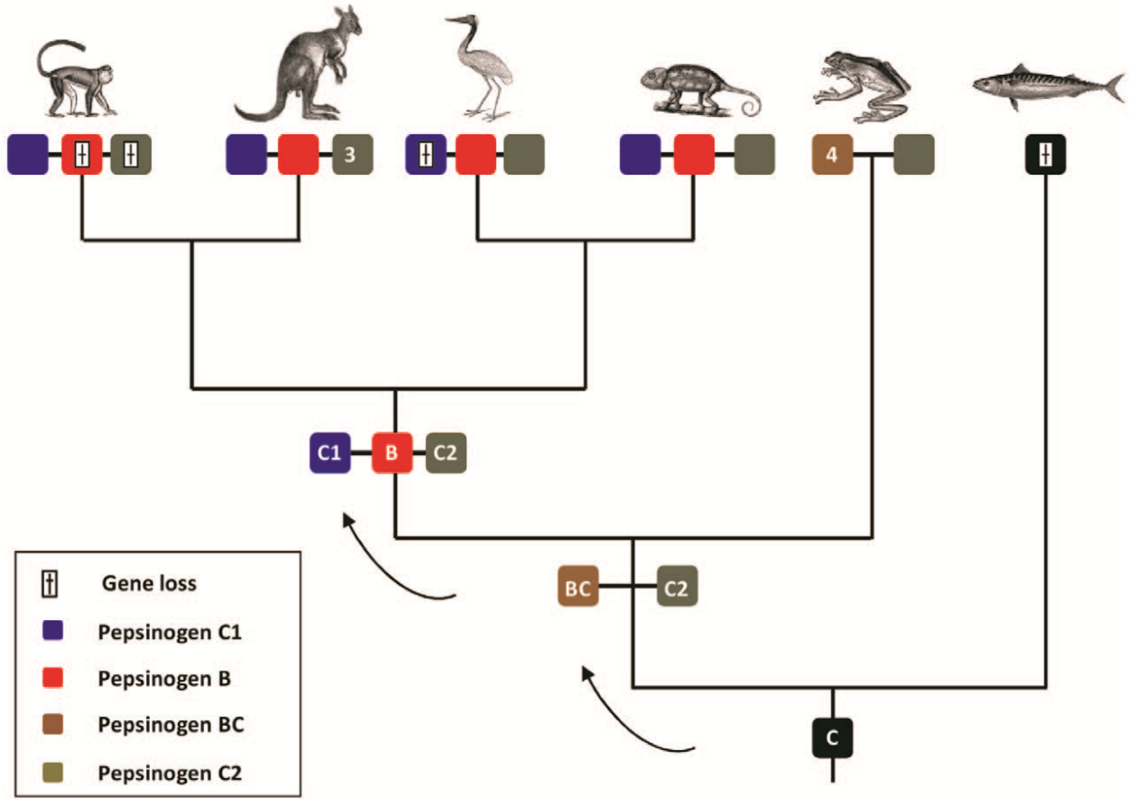


**Figure 5. Proposed evolutionary history and duplication timings of the *Pgc* gene family in vertebrates.** Numbers inside each box denotes gene numbers.
doi:10.1371/journal.pone.0032852.g005

with the multiplicity of *Pga* genes apparently linked to food habits [7]. Two fold reasons support this hypothesis: a higher number of *Pga* genes contributes to a higher level of pepsin in the stomach, and distinct isoforms (A1 and A2) have evolved distinct proteolytic specificities [7]. Coincidently, the majority of the gene expansion events observed in the *Pgc* gene family notably coincides with the invasion of terrestrial habitats. In effect, the extensive increase of *Pgc* gene lineages and independent expansions is uniquely observed in tetrapods, at two distinct moments, before the divergence of amphibians and amniotes respectively. Thus, we propose that the access to new dietary protein sources acted as the driving force for *Pgc* retention and functional diversification after gene duplication. Conversely, the targeted loss of some isoforms, such as the *Pgb* in some mammalian species or *Pgc1* in birds, once more resulted from changes in protein sources which rendered the retention of some *Pgc* isoforms less important.

### Conclusions

The data presented here significantly modifies our knowledge about the overall evolutionary history of the *Pgc* gene family considered so far. We show that *Pgc* has undergone episodes of expansion, loss and retention. We conclude that tandem duplications have modelled the history of *Pgc* genes, probably underscoring different enzymatic requirements and specificities towards protein dietary sources. Future experimental assays should take into account the evolutionary history and diversity of *Pgc* genes in vertebrates.

## Supporting Information

**Figure S1** ***Anolis carolinensis Pgb-like* pseudogene.** Grey and white shading indicate exon boundaries. In panel **A** *Anolis carolinensis* pseudogene PGB gene cDNA (Gene ID: 100567523). Highlighted in red we find the insertion of a guanine producing a premature stop codon downstream also in red. In Panel **B** Translation *Anolis carolinensis* pseudogene PGB gene cDNA, asterisk indicates stop codon. A frameshift mutation upstream results in a premature stop codon observed in exon 8. In Panel **C**, we provide a schematic representation of the *Anolis carolinensis* pseudo gene organization. Below, in detail a partial sequence alignment of exon 8 from *Anolis carolinensis* PGC1 (Ac-PGC1) *Anolis carolinensis pseudogene* (Ac-PGBΨ) and *Homo sapiens* PGC1 (Hs-PGC1). Highlighted in red frame shift mutation caused by the insertion of an guanine leading to a premature stop codon downstream also highlighted in red. Panel **D** NJ tree showing that the *AcPgb* pseudogene robustly groups with the *Pgb* orthologue.
(PPT)

**Figure S2 Structural analysis of the PGC sequences suggests distinct substrate specificities.** Hs - *Homo sapiens*; Md - *Monodelphis domestica*; Ss - *Sus scrofa*; Ac - *Anolis carolinensis*; Xt - *Xenopus tropicalis*; Ga - *Gasterosteus aculeatus*, To - *Thunnus orientalis*. All pepsinogen 3D theoretical models present a bilobal structure with the substrate binding cleft located in the middle of the two lobes (Panel A small image, 1- N-terminal; 2-substrate binding cleft and 3- C-terminal). Red corresponds to the location of the S1 subsite residues. Dark green corresponds to the S1′subite and lime green corresponds to the Asp32 and Asp217 residues. Models show a highly similar 3D structure within each PGC group (e.g. PGC1) in contrast, when comparing between groups (e.g. PGC1 and PGC2) it is possible to detect subtle differences in the enzyme structure, such as location of the S1 and S1′subsites, exposure of the active aspartic residues and in the general architecture of the binding cleft. In Panel (A) the hsPGC corresponds to the 1HTR crystal structure available at Protein Database (PDB), and which is highly similar to other PGC models presented, at position 7 we observe a methionine that impacts the cleft structure and is located near Asp32. In panel (B) three models of PGB are presented, at the equivalent position these models present an Isoleucine or and Phenylalanine which are bulky hydrophobic residues that may contribute to the narrowing of the cleft. In panel (C) PGBC models also present a subtle enlargement of the cleft possibly due to the distinct orientation of the methionine residue at position 7. In panel (D) PGC2 models show that the catalytic aspartic residues are more exposed in comparison to other PGC proteins and these models also present a larger cleft possibly due to an alternative Leucine residue at position 7. In panel (E) we observe that fish PGC models present a small N-terminal region in comparison with the other models. It is possible to observe that the subite S1 is located further from the active site in comparison with the other models, finally due to a deletion in the sequence fish PGC present an non hydrophobic asparagine residue at position 7 opposing the hydrophobic residues encountered at this location, this is a comparatively small residue possibly leading to an enlargement of substrate binding cleft in this region.
(TIF)

## Author Contributions

Conceived and designed the experiments: LFC ML-M JMW. Performed the experiments: LFC ML-M JMW. Analyzed the data: LFC ML-M OMG JMW. Contributed reagents/materials/analysis tools: LFC ML-M OMG JMW. Wrote the paper: LFC ML-M JMW.

## References

1. Richter C, Tanaka T, Yada RY (1998) Mechanism of activation of the gastric aspartic proteinases: pepsinogen, progastricsin and prochymosin. Biochem J 335: 481–490.
2. Foltmann B (1981) Gastric proteinases-structure, function, evolution and mechanism of action. Essays Biochem. pp 52–84.
3. Kageyama T (2002) Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. Cellular and Molecular Life Sciences 59: 288–306.
4. Borrelli L, De Stasio R, Filosa S, Parisi E, Riggio M, et al. (2006) Evolutionary fate of duplicate genes encoding aspartic proteinases. Nothepsin case study. Gene 368: 101–109.
5. Carginale V, Trinchella F, Capasso C, Scudiero R, Riggio M, et al. (2004) Adaptive evolution and functional divergence of pepsin gene family. Gene 333: 81–90.
6. Ordoñez G, Hillier L, Warren W, Grützner F, López-Otín C, et al. (2008) Loss of genes implicated in gastric function during platypus evolution. Genome Biology 9: 1–11.
7. Narita Y, Oda S, Takenaka O, Kageyama T (2010) Lineage-Specific Duplication and Loss of Pepsinogen Genes in Hominoid Evolution. Journal of Molecular Evolution 70: 313–324.
8. Narita Y, Oda S, Kageyama T (2006) Rodent monophyly deduced from the unique gastric proteinase constitution and molecular phylogenetic analyses using pepsinogen-C cDNA sequences. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 1: 273–282.
9. Tanji M, Yakabe E, Kubota K, Kageyama T, Ichinose M, et al. (2009) Structural and phylogenetic comparison of three pepsinogens from Pacific bluefin tuna: Molecular evolution of fish pepsinogens. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 152: 9–19.
10. Ikuzawa M, Inokuchi T, Kobayashi K, Yasumasu S (2001) Amphibian Pepsinogens: Purification and Characterization of Xenopus Pepsinogens, and Molecular Cloning of Xenopus and Bullfrog Pepsinogens. Journal of Biochemistry 129: 147–153.
11. Sakamoto N, Saiga H, Yasugi S (1998) Analysis of Temporal Expression Pattern and cis-Regulatory Sequences of Chicken Pepsinogen A and C. Biochemical and Biophysical Research Communications 250: 420–424.
12. Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, et al. (2011) Geneious v5.4.
13. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood,

Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution.

14. Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics. pp 2947–2948.
15. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406–425.
16. Abascal F, Zardoya R, Posada D (2005) ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21: 2104–2105.
17. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.
18. Moore SA, Sielecki AR, Chernaia MM, Tarasova NI, James MNG (1995) Crystal and Molecular Structures of Human Progastricsin at 1.62 Å resolution. Journal of Molecular Biology 247: 466–485.
19. Sali A, Blundell TL (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 234: 779–815.
20. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951–960.
21. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research 33: W244–W248.
22. Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods in enzymology 277: 396–404.
23. Schrödinger L The PyMOL Molecular Graphics System Version 1.3.
24. Narita Y, Oda S, Moriyama A, Kageyama T (2002) Primary structure, unique enzymatic properties, and molecular evolution of pepsinogen B and pepsin B. Archives of Biochemistry and Biophysics 404: 177–185.
25. Kageyama T (2006) Roles of Tyr13 and Phe219 in the Unique Substrate Specificity of Pepsin B. Biochemistry 45: 14415–14426.
26. Galea CA, Dalrymple BP, Kuypers R, Blakeley R (2000) Modification of the substrate specificity of porcine pepsin for the enzymatic production of bovine hide gelatin. Protein Science 9: 1947–1959.
27. Shintani T, Nomura K, Ichishima E (1997) Engineering of Porcine Pepsin; Alteration of S1 substrate specificity of pepsin to those of fungal aspartic proteases by site directed mutagenesis. Journal of Biological Chemistry 272: 18855–18861.
28. Zhang Y, Li H, Wu H, Don Y, Liu N, et al. (1997) Functional implications of disulfide bond, Cys45–Cys50, in recombinant prochymosin. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1343: 278–286.
29. Hartsuck JA, Koelsch G, Remington SJ (1992) The high-resolution crystal structure of porcine pepsinogen. Proteins: Structure, Function, and Bioinformatics 13: 1–25.
30. Hassan MI, Toor A, Ahmad F (2010) Progastriscin: Structure, Function, and Its Role in Tumor Progression. Journal of Molecular Cell Biology 2: 118–127.
31. Auer H, Glick D (1986) The mechanism of activation of porcine pepsinogen. Nature. pp 664–664.
32. Khan AR, Cherney MM, Tarasova NI, James MNG (1997) Structural characterization of activation intermediate 2 on the pathway to human gastricsin. Nat Struct Mol Biol 4: 1010–1015.
33. Kurokawa T, Uji S, Suzuki T (2005) Identification of pepsinogen gene in the genome of stomachless fish, Takifugu rubripes. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 140: 133–140.
34. Kuraku S, Kuratani S (2011) Genome-Wide Detection of Gene Extinction in Early Mammalian Evolution. Genome Biology and Evolution 3: 1449–1462.
35. Cañestro C, Catchen JM, Rodriguez-Mari A, Yokoi H, Postlethwait JH (2009) Consequences of Lineage-Specific Gene Loss on Functional Evolution of Surviving Paralogs: ALDH1A and Retinoic Acid Signaling in Vertebrate Genomes. PLoS Genet 5: e1000496.
36. Lynch M, Force A (2000) The Probability of Duplicate Gene Preservation by Subfunctionalization. Genetics 154: 459–473.
37. Lynch M, O'Hely M, Walsh B, Force A (2001) The Probability of Preservation of a Newly Arisen Gene Duplicate. Genetics 159: 1789–1804.

## SUPPLEMENTARY MATERIAL

### SUPPORTING FIGURE 1

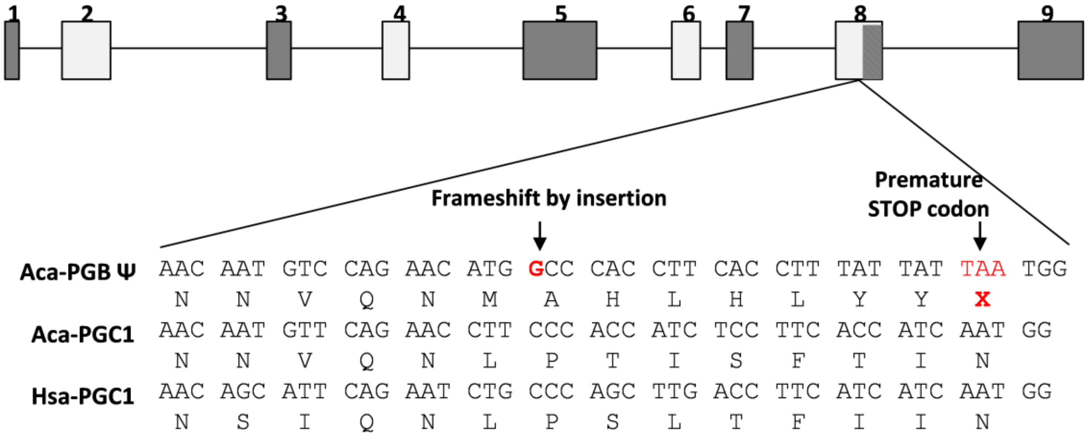
**A**

```
ATGAAGTGGTTGATCCTTTTCTTGGTTTGTCTCCACCTGTCAGAGGGACTGGAGAGAGTCGTTCTGAAGAAAG
GGAAATCCATTCGAGAGAACATGAAAGAGAAAGGTGTGCTGGAGGAATTTTTGAAGAACAACCGTGTTGATCC
TGCATTGAAATACCACTTTAATGAATACAATGTAGCTTATGAACCAATTAGCAATAACTTAAATTCTTTCTAC
TTTGGAGAGATTAGTATTGGGACACCACCACAGAATTTCCTGGTTCTTATGGATACTGGCTCTGCCAATCTTT
GGGTGCCATCTGTGTATTGCAATACTGCTGCATGTGGCAACCACAACAGATTCAACCCGAGTGCATCCTCTAC
ATACACTAACAATGGACAGACTTTTAGCCTGTACTATGGAAGTGGCAGCCTAACTGTTATGCTAGGGTATGAT
ACTGTGCAGGTCCAGAACATTGTTGTACGCAACCAGGAAATTGGCCTCAGTCAGAATGAACCTTCCAGTCCTT
TCTATTATGCCAGTTTTGATGGTATTTTGGGGATGTCTTATCCTTCTGCAGCTGTAGGCCATGTAGGGGGTTA
CACTATTATGCAGCAGATGCTGAGGCAAGGCCAGCTCTCTGAACCCATCTTCAGCTTCTATTTCTCTCGACAA
CCAACTGCTCAGTATGGAGGAGAATTGATCTTGGGAGGTATTGACACCCAGATGTTCTCTGGGGAAATTACCT
GGGCGCCTGTCACCCGTGAGGCTTACTGGCAAATTGGAATTGAGGAATTTCTAATTGGCAACCAGGCTACTGG
CTTGTGTAGCCAAGGCTGTCAGGCAGTTGTGGATACTGGGACGTATCTGCTGGCAGTGCCACAACAGTACATG
AGTACTTTCCTACAAGCTGTGGGAGGCCAAGAATATAATGGTGAGTATGAGGTGAACAATGTCCAGAACATGG
CCCACCTTCACCTTTATTATTAATGGATCCCAGTTCCCACTACCACCTTCTGCCTACATTGCCAAATGTGAGT
ATATAAGAAACAATGGCTATTGTACAGTTCAGATTGAGGCCACATACTTGCCTTCCCCAACTGGAGAGCCATT
ATGGATCTTTGGTGACGTCTTCCTCAAGGAGTATTATTCCGTCTACGATATGGCCAACAACAGGGTGGGCTTT
GCACCATCAGCTTAA
```

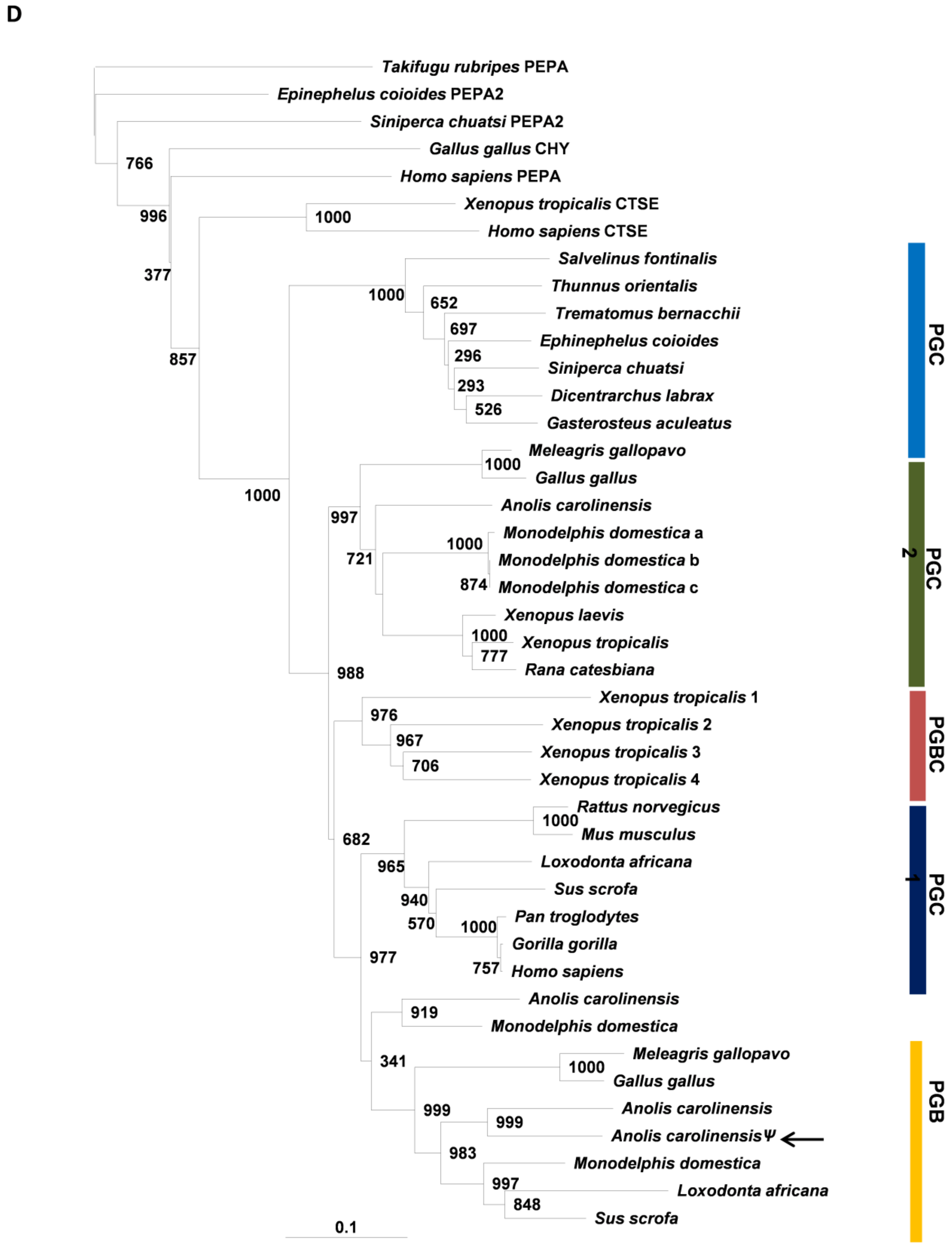
**B**

```
MKWLILFLVCLHLSEGLERVVLKKGKSIRENMKEKGVLEEFLKNNRVDPALKYHFNEYNVAYEPISNNLNSFYF
GEISIGTPPQNFLVLMDTGSANLWVPSVYCNTAACGNHNRFNPSASSTYTNNGQTFSLYYGSGSLTVMLGYDTV
QVQNIVVRNQEIGLSQNEPSSPFYYASFDGILGMSYPSAAVGHVGGYTIMQQMLRQGQLSEPIFSFYFSRQPTA
QYGGELILGGIDTQMFSGEITWAPVTREAYWQIGIEEFLIGNQATGLCSQGCQAVVDTGTYLLAVPQQYMSTFL
QAVGGQEYNGEYEVNNVQNMAHLHLYY*wipvptttfclhcqm*NNGYCTVQIEATYLPSPTGEPLWIFGDVFL
KEYYSVYDMANNRVGFAPSA*
```

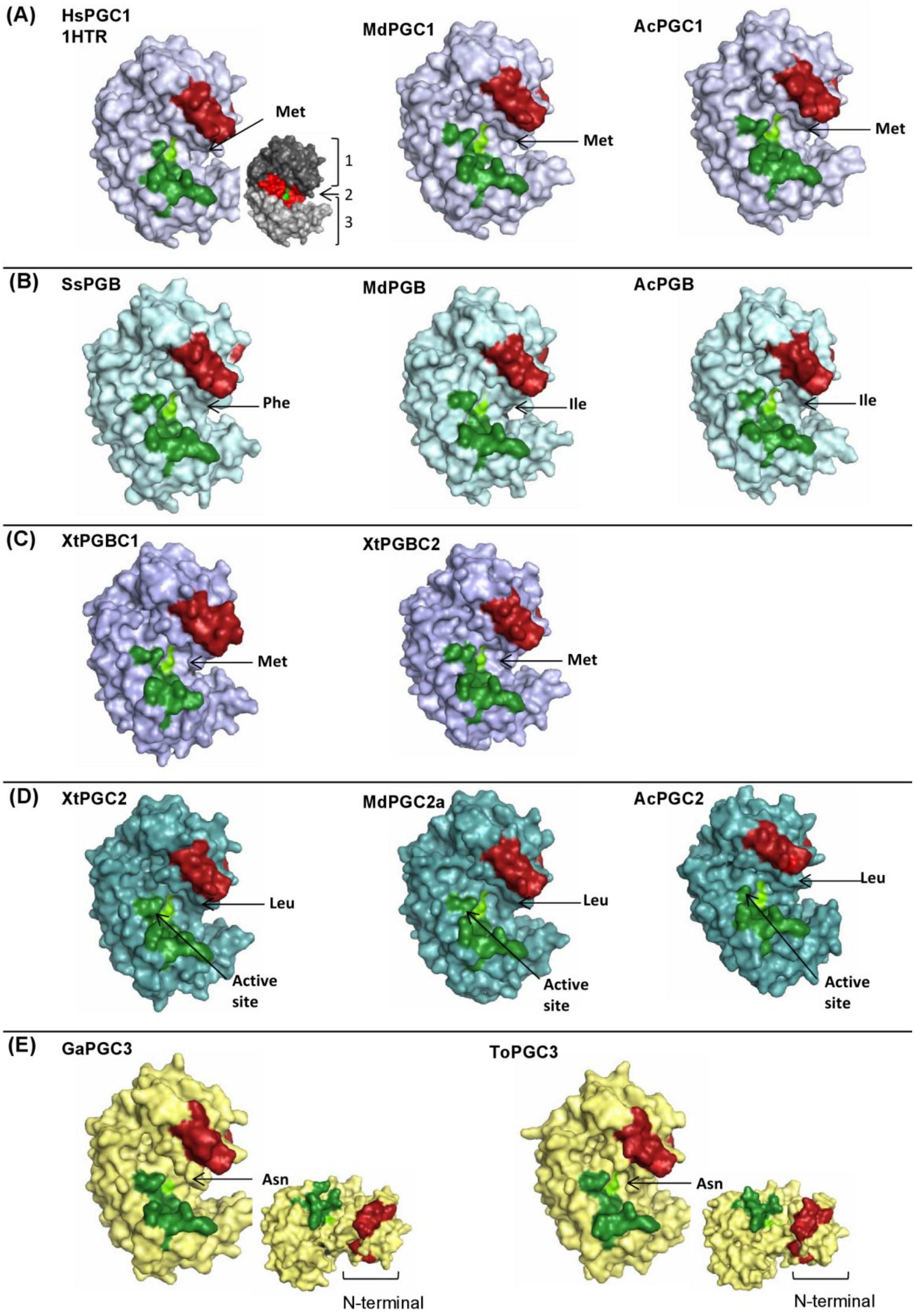
**C**



```
Aca-PGB Ψ   AAC AAT GTC CAG AAC ATG GCC CAC CTT CAC CTT TAT TAT TAA TGG
             N   N   V   Q   N   M   A   H   L   H   L   Y   Y   X
Aca-PGC1    AAC AAT GTT CAG AAC CTT CCC ACC ATC TCC TTC ACC ATC AAT GG
             N   N   V   Q   N   L   P   T   I   S   F   T   I   N
Hsa-PGC1    AAC AGC ATT CAG AAT CTG CCC AGC TTG ACC TTC ATC ATC AAT GG
             N   S   I   Q   N   L   P   S   L   T   F   I   I   N
```


D
Takifugu rubripes PEPA
Epinephelus coioides PEPA2
Siniperca chuatsi PEPA2
766
Gallus gallus CHY
Homo sapiens PEPA
996
Xenopus tropicalis CTSE
1000
Homo sapiens CTSE
377
Salvelinus fontinalis
Thunnus orientalis
1000
652
Trematomus bernacchii
697
Ephinephelus coioides
857
296
Siniperca chuatsi
293
Dicentrarchus labrax
526
Gasterosteus aculeatus
Meleagris gallopavo
1000
Gallus gallus
1000
Anolis carolinensis
997
Monodelphis domestica a
1000
721
Monodelphis domestica b
874
Monodelphis domestica c
Xenopus laevis
1000
Xenopus tropicalis
777
988
Rana catesbiana
Xenopus tropicalis 1
976
Xenopus tropicalis 2
967
Xenopus tropicalis 3
706
Xenopus tropicalis 4
Rattus norvegicus
1000
682
Mus musculus
965
Loxodonta africana
Sus scrofa
940
Pan troglodytes
570
1000
Gorilla gorilla
977
757
Homo sapiens
Anolis carolinensis
919
Monodelphis domestica
341
Meleagris gallopavo
1000
Gallus gallus
999
Anolis carolinensis
999
Anolis carolinensisΨ
983
Monodelphis domestica
997
Loxodonta africana
848
Sus scrofa
0.1
PGC
PGC 2
PGBC
PGC 1
PGB

## SUPPORTING FIGURE 2



## V.2 Unusual Loss of Chymosin in Mammalian Lineages Parallels Neonatal Immune Transfer Strategies

Mónica Lopes-Marques, Raquel Ruivo, Elza Fonseca, Ana Teixeira, L. Filipe C. Castro

**Under Revision**

# UNUSUAL LOSS OF CHYMOSIN IN MAMMALIAN LINEAGES PARALLELS NEONATAL IMMUNE TRANSFER STRATEGIES

Mónica Lopes-Marques[1,2], Raquel Ruivo[1], Elza Fonseca[1,3], Ana Teixeira[1], and L. Filipe C. Castro[1,3]*

[1]**CIIMAR** – Interdisciplinary Centre of Marine and Environmental Research, U. Porto – University of Porto, Porto, Portugal

[2]**ICBAS** - Institute of Biomedical Sciences Abel Salazar, U. Porto - University of Porto, Portugal

[3]Department of Biology, Faculty of Sciences, U. Porto - University of Porto, Portugal

*Corresponding author at: CIIMAR/CIMAR, U. Porto, Rua dos Bragas 289, P 4050-123, Porto, Portugal. Tel.: +351 223 401 831

**KEYWORDS:** Prochymosin; pseudogene; mammals; Immunoglobulin G; passive transfer

## ABSTRACT

Gene duplication and loss are powerful drivers of evolutionary change. The role of loss in phenotypic diversification is notably illustrated by the enzymatic repertoire involved in vertebrate digestion. Among these we find the pepsin family of aspartic proteinases, including chymosin. Previous studies demonstrated that chymosin, a neonatal digestive pepsin, is inactivated in some primates, including humans. This pseudogenization event was hypothesized to result from the acquisition of maternal immune IgG transfer. By investigating 94 mammalian subgenomes we reveal an unprecedented level of *cmy* erosion in placental mammals, with numerous independent events of gene loss taking place in primates, dermoptera, rodentia, cetacea and perissodactyla. Our findings strongly suggest that the recurrent inactivation of *cmy* correlates with the evolution of the passive transfer of IgG and uncover a noteworthy case of evolutionary cross-talk between the digestive and the immune system, modulated by gene loss.

## 1. INTRODUCTION

Gene loss has long been considered a secondary driver in adaptive evolution. Yet, the current paradigm is shifting and gene loss emerging as a pivotal player in the sculpting of evolutionary change (Albalat and Canestro, 2016). The genetic repertoire of gastric genes across vertebrate lineages, for instance, provides a remarkable example on the decisive role of gene loss in adaptive phenotypic variation: with several cases of gene expansion and gene loss with morpho-functional consequences (Castro et al., 2014; Castro et al., 2012; Kageyama, 2002; Ordoñez et al., 2008). A subset of digestive enzymes includes the pepsin family of aspartic proteinases (Pearl and Blundell, 1984). In mammals, the pepsin family consists of 5 members highly expressed in the gastric mucosa, grouped according to phylogenetics and substrate specificity: chymosin (*Cmy*), pepsin A (*PgA*), B (*PgB*), C (*PgC*), and F (*PgF*) (Carginale et al., 2004; Kageyama, 2002; Wu et al., 2009; Yakabe et al., 1991). The pepsin gene family is widely disseminated, yet erratically distributed, within the tetrapod lineage, with cases of gene expansion, pseudogenization and loss (Castro et al., 2014; Castro et al., 2012; Kageyama, 2002). For example, while *Homo sapiens* presents 3 copies of *PgA*, a single copy is found in *Anolis carolinesis* and *Xenopus tropicalis*; on the other hand, *PgA* is pseudogenized in *Monodelphis domestica* and *Mus musculus* exhibits no genomic evidence of this gene (Castro et al., 2014; Castro et al., 2012; Narita et al., 2010; Ordoñez et al., 2008). This species-specific distribution has been suggested to result from dietary adaptations; generally, higher levels of pepsinogens are found in the gastric mucosa of animals with an herbivorous diet, in contrast to omnivorous and carnivorous species (Kageyama, 2002).

In contrast to the other members of the pepsin family little is known about the evolutionary history, distribution and function of *cmy* in the mammalian lineages. Despite exhibiting a conserved quaternary structure and catalytic residues, *cmy* displays an unusual profile: low general proteolytic activity and high specificity towards milk κ-casein (Kageyama, 2002; Pearl and Blundell, 1984). Milk κ-casein, along with α- and β-caseins, belong to the secretory calcium-binding phosphoprotein gene family and provide nutritional calcium, amino acids, as well as other bioactive peptides, with putative antimicrobial activity (Caroli et al., 2009; Kawasaki and Weiss, 2003).

Additionally, caseins form heterogeneous micellar structures with κ-casein coats for increased stability (Mercier et al., 1976). Cleavage by *cmy* splits κ-casein into an insoluble para-κ-casein and a soluble caseinomacropeptide, leading to the disruption of the micelles, release of the entrapped content, and to the clotting of milk, a feature widely used in the manufacturing of dairy products (Caroli et al., 2009; Langholm Jensen et al., 2013; Mercier et al., 1976; Palmer et al., 2010). In fact, the use of *cmy* in the manufacturing of cheese is considered to be one of the earliest biological applications of enzymes, with remains of cheese found in Egyptian pots dating to approximately 3000-2800BCE (Palmer et al., 2010; Szecsi, 1992). Thus, research on *cmy* has focused mainly on the characterization of biochemical, structural and functional properties for industrial purposes.

Although *cmy* activity is directly related to milk clotting, it has been indirectly correlated with colostrum-dependent immunoglobulin transfer (Borghesi et al., 2014; Hurley and Theil, 2011; Kageyama, 2002). Several key aspects were suggested to enhance immunoglobulin transfer from colostrum. First, the mild enzymatic landscape of the neonate digestive tract, which protects immunoglobulins from proteolysis (Bela Szecsi and Harboe, 2013; Foltmann, 1992);  and, the presence of trypsin inhibitors in the colostrum, conferring added protection against enzymatic digestion in the small intestine (Foltmann, 1992). Thus, in colostrum dependent IgG transfer, *cmy* would contribute to the mild environment while allowing the release of micellar immunoglobulins and whey proteins (Bela Szecsi and Harboe, 2013; Foltmann, 1992). However, IgG transfer strategies vary across mammals. In humans, for instance, IgG is transferred from mother to fetus during the last stages of gestation. Curiously, *cmy* was found to be a pseudogene in humans, due to a shift in the reading frame, and consequent premature stop codon, caused by a deletion in exon 4 (Ord T, 1990). Despite the pseudogene status in human, *cmy* can be found in several tetrapod species (mammals, reptiles and bird (Ordoñez et al., 2008)). A cryptic orthologue was also suggested in teleost genomes, on the basis of a conserved orthologous syntenic region. However, contrary to the other members of the pepsin family with inconsistent gene distribution in tetrapods, the evolutionary history and distribution of *cmy* is largely unknown. Thus, the emerging questions are (1) whether the pseudogenized condition detected in humans is unique, or conversely, if it represents a wider genomic trait of mammals, and, if so, (2) does *cmy* pseudogenization follows the acquisition of maternal immune transfer strategies. Here, we sought to illuminate the evolutionary history of the *cmy* gene and its correlation with feeding and immune transfer strategies by providing an extensive analysis of available mammalian genomes.

## 2. MATERIALS AND METHODS

### 2.1 SEQUENCE ANALYSIS

All major mammalian lineages with available genome data in Ensembl and GenBank were searched with blastp and blastn using as query *Bos taurus Cmy* amino acid and/or nucleotide sequence. *Cmy*-like nucleotide sequences were retrieved for the following lineages: Monotremata; Marsupialia; Cingulata, Trubulidentata, Macroscidea, Afroscidea, Proboscidea, Sirenia, Eulipotyphyla, Chiroptera, Perrissodactyla, Carnivora, Cetacea, Artiodactyla, Lagomorpha, Hystricomorpha, Sciuromorpha, Myomorpha, Scandentia, Dermoptera, Strepsirhini, Haplorrhine, Platyrrhini, Cercopthecoidea, Hominidae. Non-mammalian lineages namely reptiles and birds were also searched and the corresponding nucleotide sequences retrieved. A total of 99 *Cmy*-like nucleotide sequences recovered and corresponding accession files were inspected to determine if the annotated RefSeq transcript was modified relative to its source genomic, if affirmative the corresponding genomic region of the *Cmy-like* gene would be retrieved, and further examined to determine coding status (Supplementary Table1).

### 2.2 GENE ANNOTATION AND MUTATIONAL VALIDATION

Using *Bos taurus* prochymosin nucleotide sequence (NM_180994.2) as reference, each exon was isolated and mapped to the genomic region of the candidate pseudogenes using Geneious V7.1.9 map to reference tool. The aligned regions were individually screened for ORF disrupting mutations (frameshift and premature stop codons) and then concatenated to obtain a predicted cDNA**.** Validation of the identified ORF abolishing mutations was performed by blastn searches in available sequence read archive (SRA) and Trace Archive in NCBI (when available) using as query the nucleotide sequence of the exon containing the mutation. Blast hits were uploaded to Geneious V7.1.9 and mapped to the corresponding exon, the final alignment with SRA reads was inspected to remove poorly aligned sequences and to confirm mutation status. The validation of at least one abolishing mutation per species by SRA reads and or Trace archives reads was performed.

### 2.3 PHYLOGENY AND SELECTION ANALYSIS

Initial screening and gene annotation identified 30 potential pseudogenes. The remaining 69 coding *CMY* ORFs were selected for phylogenetic analysis. An initial sequence alignment was performed to identify and purge partial sequences from further analysis. Nucleotide sequence alignment for phylogenetic analysis was performed in MAFFT (Katoh et al., 2005; Katoh and Toh, 2008) with L-INS-I method. The resulting sequence alignment was stripped of all columns containing gaps leaving a total of 1072 positions for phylogenetic analysis. Maximum likelihood phylogenetic analysis was performed in PhyML V3.0 (Guindon et al., 2010) and the evolutionary model was determined using the smart model selection (SMS) option resulting in a GTR +G+I+F. The branch support was calculated using aBayes. The resulting tree was analysed in Fig Tree V1.3.1 available at http://tree.bio.ed.ac.uk/software/figtree/ and rooted on the bird and reptile clade.

The analysis of the selective regime was performed exclusively in the mammalian sequences. These were aligned by codon translation in Geneious V7.1.9; exon1 was stripped from all sequences, as well as, columns containing 90% of gaps. The final sequence alignment was submitted to the Datamonkey Webserver Suite (Pond and Frost, 2005; Pond et al., 2005) and selective strength was calculated using RELAX (Wertheim et al., 2014). Data type was set to codon and the genetic code was set to universal. For each clade analysed one analysis was run in RELAX were the target clade was set as test branch while the remaining clades remained as reference branches. Time tree containing all species for overall analysis was created by submitting the list of analysed species to the time tree public knowledge-base (Hedges et al., 2006; Hedges et al., 2015; Kumar and Hedges, 2011).

## 3. RESULTS

### 3.1 SEQUENCE ANALYSIS AND GENE ANNOTATION

A total of 94 species covering all major mammalian lineages were examined for the presence of *cmy-like* sequences. For each retrieved sequence the corresponding Gene Bank file was inspected to determine the gene coding status. Sequence search and analysis returned a total of 30 candidate pseudogenes and no annotation of a *cmy-like* sequence was found in 4 species: *Heterocephalus glaber* (naked mole-rat); *Octodon degus* (Degu), *Fukomys damarensis* (Damaraland mole-rat) and *Chinchilla lanigera* (common chinchilla).

All 30-candidate pseudogenes were individually inspected and re-annotated using the corresponding species-specific genomic data. The analysis of the *H. sapiens* CMYP (OMIM#118943) genomic region, confirmed a frameshift in exon 4 produced by a 1bp deletion, followed by a termination codon in exon 5, and a second frameshift in exon 6 produced by a 2bp deletion. All of these observations are concurring with previous studies (Fig. 1) (Kolmer et al., 1991; Örd et al., 1990) and validate our annotation strategy. This method was subsequently applied in the analysis of the genomic regions of the identified candidate *cmy* pseudogenes revealing several ORF-abolishing mutations in all of the hominoidea species analysed (Fig. 1). ORF-disrupting mutations were also found in the sister clade, cercopithecoidea, where we found no remnants of exon 3 and exon 4 in the corresponding genomic region suggesting marked *cmy* erosion by the complete deletion of these exons in these species, in addition to several observed ORF abolishing mutations (Fig.1). Interestingly, a cross species analysis in hominoidea and cercopithecoidea uncovers a conserved single mutation that spans throughout all analysed *cmy*-like sequences, namely a termination codon in exon 5. This mutation was further corroborated with SRA reads and, when available, Trace archives sequences (Supplementary material 1). Still, within the primates an additional candidate *cmy* pseudogene was identified and confirmed by SRA reads in the new world monkey *Aotus nancymaae*; however this species shared no mutations with the previously analysed primates (Fig. 1 and Supplementary material 2).

Regarding *Galeopterus variegatus* (Sunda flying lemur) the only representative of dermoptera with available genome data, gene annotation was unable to retrieve exon 1 although this genomic region is fully sequenced up to the neighboring gene (not shown) (Fig. 1). Additionally, frameshift mutations were identified with 1 bp deletion in exon 2 and 2 bp deletion in exon 6, the latter being confirmed by SRA reads (Fig. 1 and Supplementary material 3).

In rodentia although no annotation of *cmy* gene was found in *H. glaber*, *C. porcellus, O. degus*, *F. damarensis*, and *C. lanigera*, the analysis of corresponding genomic region uncovered *cmy*-like relic sequences (Fig. 1). Gene annotation of these genomic regions revealed a high deterioration of the extant *cmy* sequence with the loss of several exons in accumulation to various disrupting mutations and possibly an event of complete gene loss for *C. lanigera, O. degus,* and *F. damarensis* . Mutations found in rodentia were further validated by SRA reads for *C. porcellus*, 4bp deletion and termination codon in exon 4, *M. auratus*, 1bp deletion in exon 4, and *H. glaber* 1 bp deletion in exon7 (supplementary material 4). In contrast, for *Marmota marmota* we found only one nucleotide substitution that resulted in a premature terminator codon in exon 7. However, we were unable to validate this mutation given that no SRA or Trace archives are available for this species.

In perissodactyla the two equidae analysed species, *Equus caballus* (Common horse) and *Equus przewalskii* (Dzungarian horse), show a conserved mutational pattern with the deletion of exon 2, 3 and 4, several indels along the gene sequence and sharing a common termination codon mutation in exon 6; while in the rhinocerotidae *Ceratotherium simum* (white rhinoceros) we found similar mutational events however not conserved with equidea. For example, *C. simum* also presents a termination codon in exon 6 however not in the same coordinates as the equidae. Frameshift mutation in exon 6 in *E. caballus* and stop codon in *C. simum* were validated by SRA reads (supplementary material 5).

Regarding cetacea the only specie in this clade presenting a candidate pseudogene was *Orcinus orca* with 1bp deletion in exon 3, confirmed by SRA reads (supplementary material 6). Finally, in the monotremata *Ornithorhynchus anatinus* (Platypus) a 5bp insertion in exon 4 was found in genomic data (Fig. 1). However, after searching all available SRA archives no *cmy*-like reads were retrieved, therefore remaining unconfirmed.

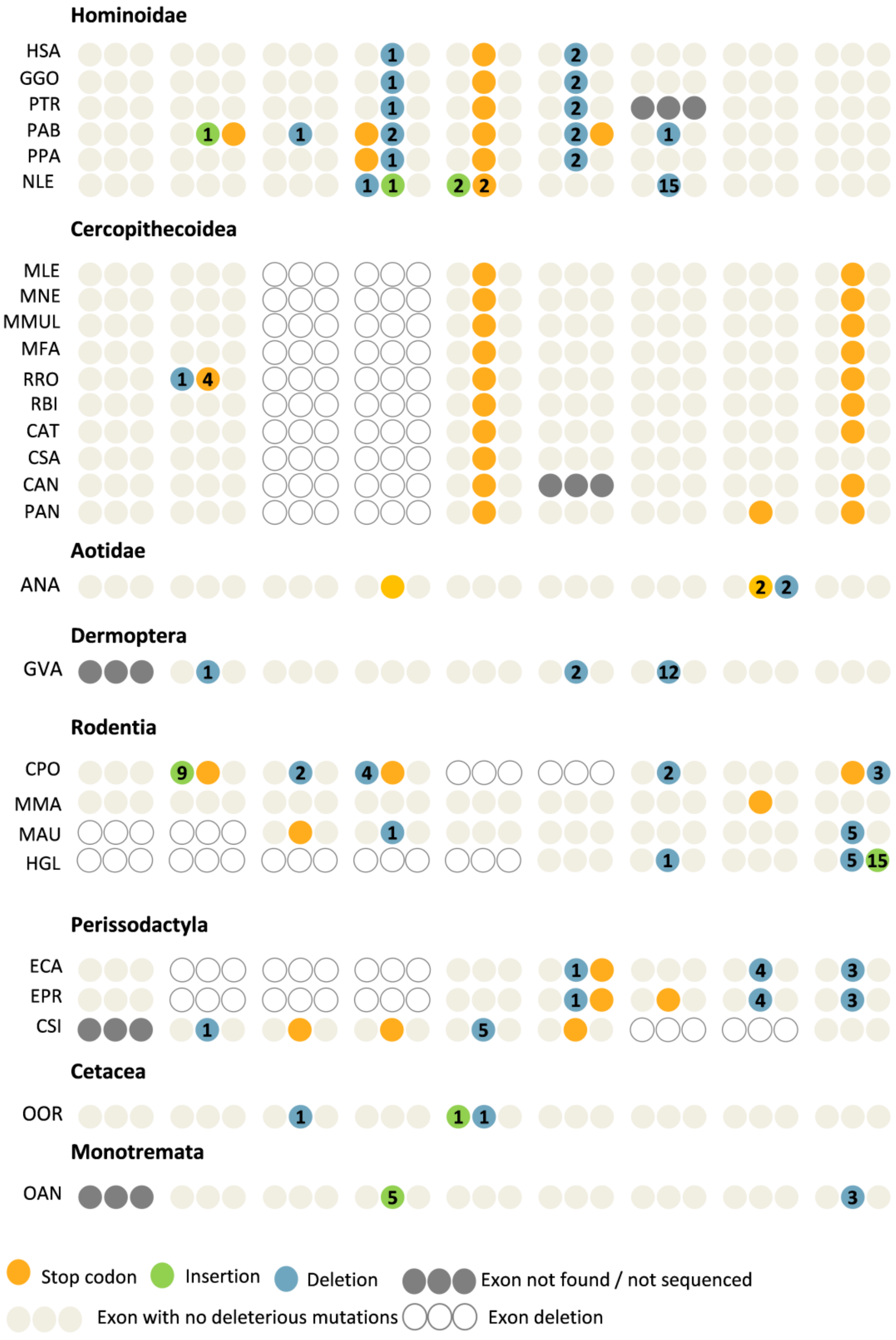


**Figure 1: Chymosin gene annotation.** A group of three circles represent a single exon, empty circles correspond to exon deletion, grey circles correspond to exon not found or located in regions with poor genome coverage, numbers in the circles correspond to the number of nucleotides inserted or deleted.

HSA- *Homo sapiens,* GGO - *Gorilla gorilla* , PTR- *Pan troglodytes* , PAB- *Pongo abelii* , PPA- *Pan paniscus*, NLE- *Nomascus leucogenys,* MLE- *Mandrillus leucophaeus* , MNE-*Macaca nemestrina* , MMUL -*Macaca mulatta* , MFA-*Macaca fascicularis* , RRO-*Rhinopithecus roxellana* ,RBI *Rhinopithecus bieti* ,CAT-*Cercocebus atys*, CSA-*Chlorocebus sabaeus*, CAN-*Colobus angolensis*, PAN-*Papio anubis,* ANA-*Aotus nancymaae*, GVA-*Galeopterus variegatus* , HGL-*Heterocephalus glaber,* CPO- *Cavia porcellusl,* MMA- *Marmota marmota,* MAU- *Mesocricetus auratus, ECA-Equus caballus*, EPR-*Equus przewalskii,* CSI- *Ceratotherium simum*, OOR-*Orcinus orca* and OAN-*Ornithorhynchus anatinus*.

The annotation of all 30 candidate pseudogenes revealed two cases of poor annotation and/or genome assembly in carnivores. Initial analysis indicated that *Ursus maritimus* (Polar bear) and *Leptonychotes weddellii* (Weddell seal) presented candidate *cmy* pseudogenes. Gene annotation revealed a 2bp frameshift in exon 9 in *L. weddellii* and 5bp frameshift in exon 7 in *U. maritimus.* However, the validation of these mutations by SRA search displayed dissimilarities between the source genomic sequences in NCBI and the SRA read archives. In the case of *L. weddellii* 3 reads from SRR332059 confirmed the two-nucleotide deletion, while 10 reads from 2 independent SRA runs SRA353375 and SRA353374 indicate otherwise. For *U. maritimus* several SRA projects available were searched yet none of the reads recovered (above 300) confirmed the 5bp insertion in exon 7 (Supplementary material 7). Thus, the *cmy* sequences for *U. maritimus* and *L. weddellii* were regarded as most probably coding.

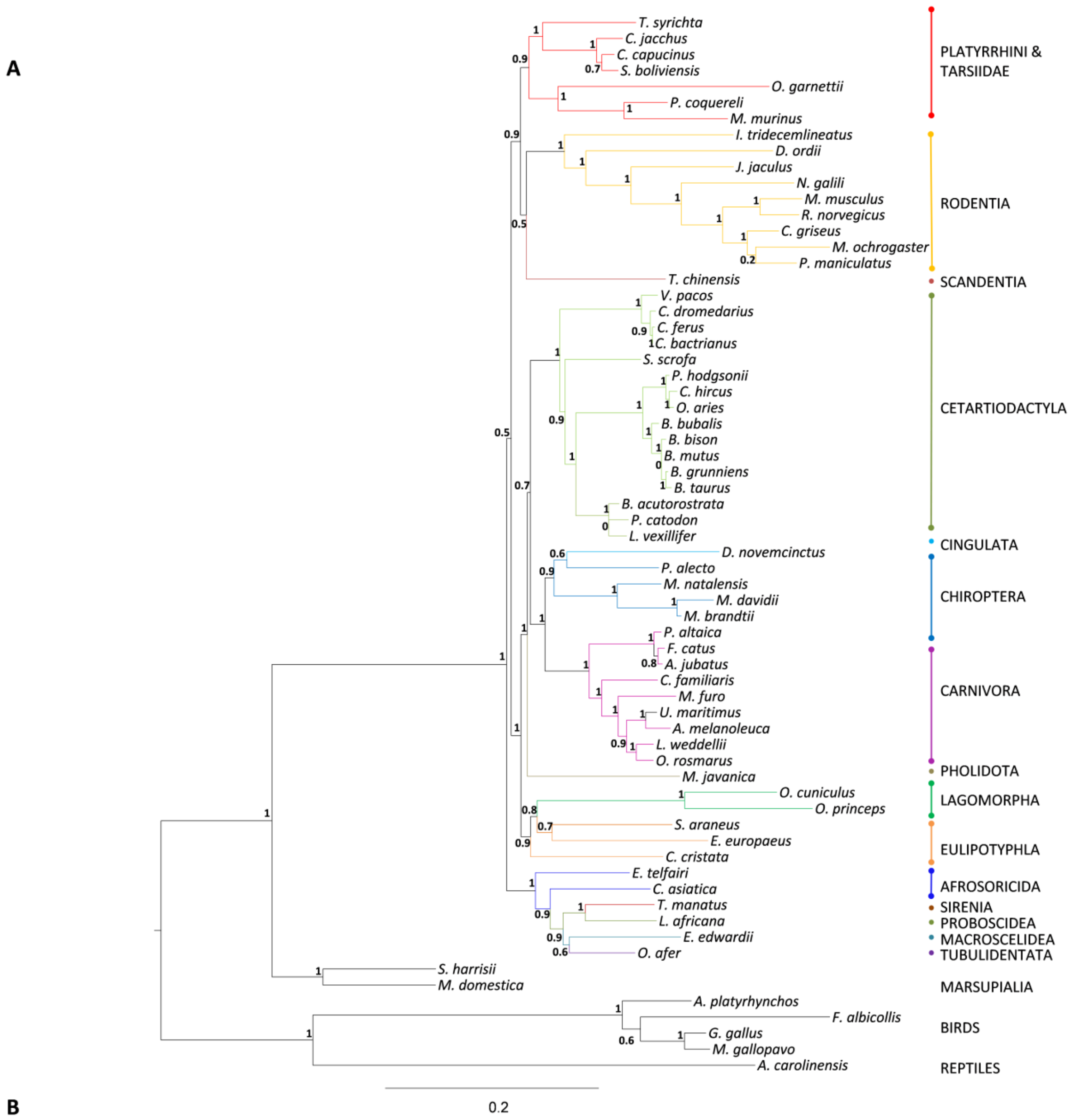
In summary, the *cmy* gene is present in all analysed members of the, carnivora, chiroptera, scandentia, cingulata, pholidota, lagomorpha, eulipotyphla, macroscelidea, afrosoricida, sirenia, proboscidea, tubulidentata and marsupialia. Inversely, in perissodactyla and dermoptera we find no coding *cmy* sequences, while in primates, rodentia, and cetacea we find a mixed profile with members showing a coding *cmy* and other species displaying strong signs of gene erosion.

### 3.2 PHYLOGENETICS AND SELECTION ANALYSIS

The initial screening and sorting of coding *cmy* genes and candidate pseudogenes, together with the purging of 3 partial coding *cmy* sequences from *Pteropus vampyrus* (large flying fox-XM_011369072.1), *Eptesicus fuscus* (big brown bat-XM_008147352.1) and *Myotis lucifugus* ( little brown bat-XM_014458936.1) left a total of 66 predicted coding sequences. Phylogenetic analysis returned an overall tree topology consistent with that of mammalian speciation: two well-supported clades one representing the placental species and one containing the marsupials, out-grouped by bird and reptile sequences (Fig.2A).

Selection analysis revealed that the platyrrhini and tarsiidae, rodents and lagomorpha branches exhibited significant relaxed selection ($k<1$) with the test branches shifting towards neutrality ($\omega=1$), while the carnivora and chiroptera branches presented a significant intensified selection ($k>1$) with the test branches shifting away from neutrality (Fig.2B).

In the platyrrhini and tarsiidae, rodents and lagomorpha groups we found that the p-values are increasingly significant as we move from lagomorpha to rodents and to platyrrhini; while for the intensified selection we show that carnivores comparatively to chiroptera present a higher significance in selection analysis. Finally, for artiodactyla, cetacea afrosoricida, sirenia, proboscidea, macroscidea and tubulidentata selection tests were not significant meaning that relaxed selection is not observed for *cmy* in these clades (Fig. 2 B).



B

| Test branch | ω Ref :ω Test | K | p-value | LR | |
|---|---|---|---|---|---|
| Playrrhini and Tarsiidae | 0,237 : 0,355 | 0.69 | 0.002 | 9,51 | Significantly Relaxed |
| Rodentia | 0,238 : 0,281 | 0.78 | 0.010 | 6,60 | Significantly Relaxed |
| Lagomorpha | 0,241 : 0,325 | 0.74 | 0.028 | 4,79 | Significantly Relaxed |
| Cetartiodactyla | 0,243 : 0,273 | 0.83 | 0.110 | 2,54 | - |
| Carnivores | 0,250 : 0,221 | 1.39 | 0.002 | 5,3 | Significantly Intensified |
| Chiroptera | 0,248 : 0,214 | 1.48 | 0.027 | 4,86 | Significantly Intensified |
| Afrosoricida, Sirenia, Proboscidea, Macroscidea and Tubulidentata | 0,255 : 0,189 | 1.31 | 0.082 | 3,01 | - |

**Figure 2:** A- Bayesian phylogenetic analysis of chymosin amino acid sequences, values at nodes indicate posterior probabilities. B. Test for relaxed selection using RELAX (Wertheim et al., 2014) in the various mammalian lineages, were K= mean selection intensity parameter, ωRef = selection rate in the background tree, ω Test= selection rate in the test branch, were p-values below 0.05 considered significant.

## 4. DISCUSSION

Here, we set out to investigate the distribution and coding status of *cmy* in mammals. The coding condition of *cmy* was previously denoted non-functional in humans (Ord T, 1990). The expanded analysis carried out in our study strongly indicates an unprecedented level of gene loss with at least 8 independent events throughout various mammalian orders (Fig 3). However, *cmy* characterization is misrepresented across the mammalian clade. Interestingly, a large body of research correlates *Cmy* with colostrum-dependent post-natal immunity transfer, namely immunoglobulin G (IgG) (Cruywagen, 1990; Foltmann, 1992; Jensen et al., 1982). In fact, high *cmy* expression in new-born mammals has been confirmed in several species: *Bos taurus* (Andrén, 1992); *Ovis aries* (Pungerčar et al., 1991); *Sus scrofa* (Foltmann et al., 1995; Foltmann et al., 1981); *Rattus norvegicus* (Kageyama et al., 2000) and *Felis catus* (Jensen et al., 1982) with expression the of *cmy* gradually ceasing during the weaning process (Kageyama, 2002). However, the mechanistic role of *cmy* in IgG transfer has been addressed almost exclusively in certatiodactyls, notably ruminants (Cruywagen, 1990). In camels, pigs and ruminants, *cmy* cleaves the abundant κ-casein, disrupts the casein micelles, and triggers the release of the entrapped whey protein phase, containing IgG, into the intestinal track for non-selective absorption within the first hours after birth (Bela Szecsi and Harboe, 2013; Cruywagen, 1990; Hurley and Theil, 2011; Mokhber-Dezfooli et al.). Besides facilitating IgG release, micelle disruption leads to the formation of a clot of insoluble para-κ-casein, which allows a slower and efficient nutrient absorption (Cruywagen, 1990; Foltmann, 1992). Curiously, the pattern of *cmy* gene loss observed in our study seems to parallel the acquisition of novel immune transfer strategies, as previously suggested for carnivores, primates and rodents (Bela Szecsi and Harboe, 2013; Jensen et al., 1982; Kageyama, 2000); yet, aside from the ruminant case study, the putative role of *cmy* in passive IgG transfer in other mammals is not fully understood.

In newborn primates, placental transfer of IgG occurs during the final stages of gestation (Coe and Lubach, 2014; Coe et al., 1994). This transfer has been found to progressively increase within the primate lineages in accordance with their evolutionary divergence pattern. In basal lineages (*O. garnettii*) a minimal prenatal immune transfer is observed at birth while in new world monkeys (*S. boliviensis)* IgG transfer is approximately 40% (Coe et al., 1994). In the old-world monkeys (*M. mulutta*, *P. troglodytes* and *H. sapiens*) full term neonates present levels of IgG that vary from approximately 75% to over 100% when compared to the progenitor (Coe et al., 1993; Coe and Lubach, 2014; Coe et al., 1994). In agreement, *cmy* pseudogenization spans from *H. sapiens* to all hominoidae and cercopthecoidea (old world monkeys), including the Japanese monkey, in accordance to the observations of Kagayema *et al* (Kageyama, 2002; Kageyama et al., 1991). Mutational analysis revealed a single founding mutation transversal to all hominoids and old world monkeys. Additionally, in old world monkeys, a second conserved mutational event was also observed. This conserved pattern suggests that *cmy* pseudogenization took place at the base of this lineage after the divergence of old world monkeys and new world monkeys, approximately 40 million years ago (Nei and Glazko, 2002; Schrago and Russo, 2003) (Fig 3). Yet, in new world monkeys (with the exception of *Aotus nancymaae*) *cmy* was surprisingly identified in the stomach of adult individuals (Kageyama, 2000). Moreover, this enzyme showed a high general proteolytic activity when compared to *cmy* from other previously characterized species (Foltmann, 1992; Kageyama, 2000). These attributes suggest that, in new world monkeys, *cmy* role was diverted to adult digestion, possibly related to their mixed insectivorous and herbivorous diet (Kageyama, 2000), again with the exception of the *Aotus nancymaae* which is known to be primarily frugivorous feeding on insects only when fruit is scarce (Baer, 1994). In accordance, basal primate lineages, with low maternal IgG transfer, exhibited an intact *cmy* gene (Fig. 4).

In rodents *cmy* gene loss was observed in one sciuromorpha (*M. marmota*), one myomorpha (*M. auratus*) and all members from the hystricomorpha suborder (Fig 3). In comparison to other mammals, the observed mutational events in rodents are possibly the most detrimental, to the extent that for several species no *cmy*-like sequence was identified in genome projects. Interestingly, once more we find a correlation between *cmy* loss and IgG prenatal transfer patterns. Rodents of the hystricomorpha suborder, present, similarly to primates, hemomonochordial placentas permitting maternal immunity transfer, namely IgGs before birth (Borghesi et al., 2014). Although, rodents from myomorfa suborder present hemotrichorial placentas, which allow for immune transfer before birth, this transfer is low and therefore is combined with post-natal passive immune transfer (Baintner, 2007; Borghesi et al., 2014; Kohl and Loo, 1984; Pentšuk and van der Laan, 2009). This dual IgG source was also observed in cats and dogs (carnivores) (Baintner, 2007).

The link between immune transfer strategies and *cmy* is also illustrated by the selective strength observed in the various clades; here we find strong evidence of relaxed selection in clades that present maternal transfer of IgG. Interestingly, relaxed selection has been previously implicated as a propeller of pseudogenization, speciation and innovation (Go et al., 2005; Wertheim et al., 2014; Wu et al., 1986). Our results suggest that relaxed selection was precursor of *cmy* pseudogenization in cercopthecoideas and hominidae, and is possibly paving the way for pseudogenization of *cmy* in rodentia and lagomorpha. In fact, previous studies were unsuccessful in isolating *cmy* from rabbit, indicating that this gene is possibly already inactivated (Kageyama and Takahashi, 1984; Kageyama et al., 1990); additionally, lagomorphs present significant immune transfer during gestation which is in agreement with the "*immune hypothesis*" (Baintner, 2007; Furukawa et al., 2014). In contrast, in platyrihinni the relaxed selection probably played an essential role in re-routing *cmy* expression from neonatal to adult stomach. In species were no maternal transfer of IgG is observed, selective force on *cmy* gene is conservative.

Surprisingly, none of the analysed perissodactyls presented a functional *cmy*. The perissodactyla order comprises three subfamilies namely rhinoceratidae, tapiridae and equidae of which genome data is only available for rhinoceratidae and equidae (Steiner and Ryder, 2011). In equidae and rhinoceratidae, several mutational events were detected, yet due to gene erosion; no founding event was identified challenging to infer the approximate timing of *cmy* loss in this lineage. Concerning immune transfer, the perissodactyla present epitheliochordial placentas that are non-permissive for immune transfer in gestation, seemingly contradicting the "*immune hypothesis*" (Baintner, 2007; Furukawa et al., 2014). However, the equidea milk presents singular characteristics: a low casein content, namely κ-casein, and high content of whey proteins, in which the main immunoglobulin is IgG in colostrum and IgA in milk (Hurley and Theil, 2011). The combination of low casein content, rapid gastric evacuation, frequent nursing and high content of whey proteins in colostrum and milk accounts for passive immune transfer in these species (Hurley and Theil, 2011; Uniacke-Lowe et al., 2010).

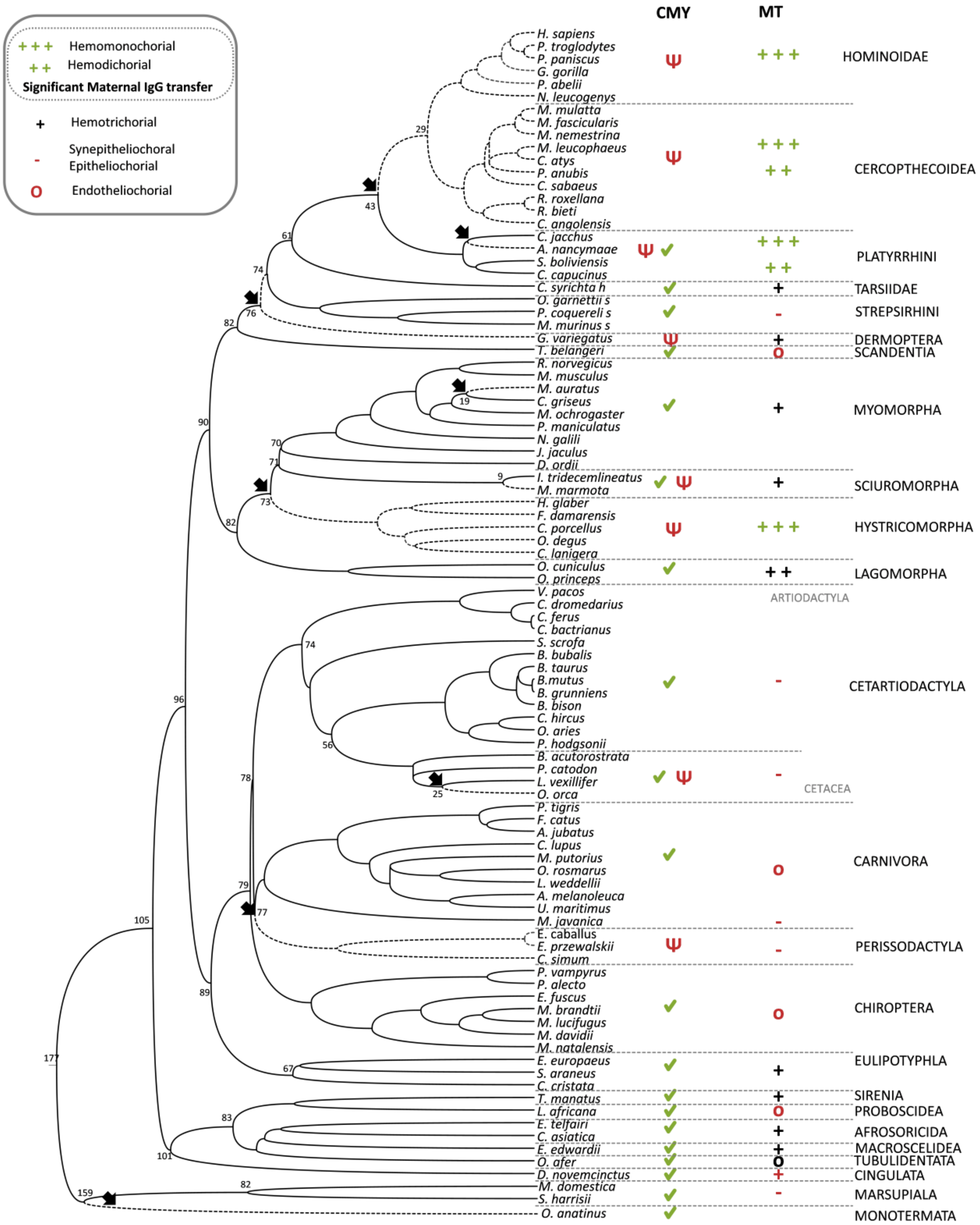


**Figure 3:** Chymosin distribution in mammals. Ψ indicates pseudogene status, √ indicates coding status, dashed lines indicated lineages were chymosin has pseudogenized, black arrows indicate independent founder events of pseudogeneization, node values correspond to divergence timings in million years ago Ma (Hedges et al., 2006; Hedges et al., 2015; Kumar and Hedges, 2011)

The remaining events of *cmy* gene loss were found in 3 different lineages were little or no information is available regarding the immune transfer process: the dermopteran *Galeopterus variegatus*, the cetacean *Orcinus orca* and the monotermata *Ornithorhynchus anatinus*. Regarding the latter, although it was not possible to validate the ORF disrupting mutations, previous studies were unsuccessful in detecting *cmy* gene expression, suggesting that platypus *cmy* is most probably pseudogenized (Ordoñez et al., 2008). If confirmed, this species presents a unique genetic makeup with no gastric proteases, in accordance with the agastric phenotype (Castro et al., 2014; Ordoñez et al., 2008).

Throughout the mammalian clade we find consistent suggestions of the role of *cmy* as a neonatal enzyme adjusted to aid the intestinal absorption of IgG. However, several questions remain unanswered. According to calf studies, *cmy* triggers a digestive cascade inducing clotting, which releases entrapped proteins and IgGs and slows-down digestion, which in turn increases absorption. Also, its specific activity towards, κ-casein, and low general proteolytic activity prevents overall degradation of IgGs. Thus, *cmy* loss could be due to shifts in colostrum composition, namely κ-casein concentration, given that α- and β-casein micelles precipitate spontaneously in the presence of calcium ions (Kawasaki and Weiss, 2003). Furthermore, the scant available data suggests that colostrum and milk compositions vary across mammals and possibly reflect distinct immune transfer strategies (Langer, 2009). For instance, a higher relative protein content was observed in colostrum from animals with exclusive post-natal transfer (certatiodactyls, including cetaceans) in comparison with mammals with maternal-only or mixed transfer profiles, such as primates, rodents, carnivores (Langer, 2009). Colostrum and milk are, in fact, complex mixtures of nutrients, digestive enzymes, activators and inhibitors (Dallas et al., 2015). Thus, it is plausible that nutritional and immune transfer strategies evolved in concert with maternal-foetus barriers and digestive tract development. In this complex scenario, *cmy* would contribute to the required digestive landscape. Curiously, IgA and IgM, predominant in human colostrum, are secreted in complex with a protective peptide (secretory complex) whereas IgG in unbound, further corroborating IgG sensitivity towards proteolytic degradation (Rojas and Apodaca, 2002).

Neonatal expression of *cmy* is not limited to mammals. In fact, *cmy* was one of the first embryonic aspartic proteases cloned in chick (Hayashi et al., 1988). During embryonic development, avian *cmy* was, also, proposed to participate in immune transfer through digestion of egg yolk protein, promoting embryonic development and immune acquisition via amniotic fluid ingestion (Foltmann, 1992; Hayashi et al., 1988; Kageyama, 2002). Despite its immune role, chicken *cmy* was suggested to hold a general proteolytic activity (Foltmann, 1992), possibly the ancestral molecular function of *cmy* re-gained in new world monkeys. Future site-directed mutagenesis analysis should uncover the contingent substitutions leading to the observed functional shifts.

## 5. CONCLUSION

To date *cmy* was reported to be pseudogenized in human and in some primates; however, our analysis reveals an unprecedented level of *cmy* loss in placental mammals, with numerous independent events of gene loss taking place, in primates, dermoptera, rodentia, cetacea and perissodactyla. Our findings strongly suggest that the recurrent erosion of *cmy* correlates with the evolution of the passive transfer of IgG, the so-called "*immune hypothesis*". Our findings uncover a noteworthy case on the role of gene loss in the evolutionary cross-talk between the digestive and the immune system and pave the way for future investigations.

## ACKNOWLEDGEMENTS

We acknowledge Fundação para a Ciência e a Tecnologia for the support to M.L.-M (SFRH/BD/84238/2012), to E.F. (SFRH/BD/79305/2011) and to R.R. (SFRH/BPD/72519/2010). LFCC research is supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and POPH - *Operational Human Potential Programme* and national funds through FCT – Foundation for Science and Technology (UID/Multi/04423/2013). We would like to thank Liliana Gonçalves for the embryo illustration in the graphical abstract.

## REFERENCES

Albalat, R., Canestro, C., 2016. Evolution by gene loss. Nat Rev Genet 17, 379-391.

Andrén, A., 1992. Production of prochymosin, pepsinogen and progastricsin, and their cellular and intracellular localization in bovine abomasal mucosa. Scandinavian Journal of Clinical and Laboratory Investigation 52, 59-64.

Baer, J.F., 1994. 5 - Husbandry and Medical Management of the Owl Monkey. Aotus: the Owl Monkey. Academic Press, San Diego, pp. 133-164.

Baintner, K., 2007. Transmission of antibodies from mother to young: Evolutionary strategies in a proteolytic environment. Veterinary Immunology and Immunopathology 117, 153-161.

Bela Szecsi, P., Harboe, M., 2013. Chymosin In: Salvesen, G. (Ed.), Handbook of Proteolytic Enzymes. Academic Press, pp. 37-42.

Borghesi, J., Mario, L., Rodrigues, M., Favaron, P., Miglino, M., 2014. Immunoglobulin Transport during Gestation in Domestic Animals and Humans—A Review. Open Journal of Animal Sciences, 323-336.

Carginale, V., Trinchella, F., Capasso, C., Scudiero, R., Riggio, M., Parisi, E., 2004. Adaptive evolution and functional divergence of pepsin gene family. Gene 333, 81-90.

Caroli, A.M., Chessa, S., Erhardt, G.J., 2009. Invited review: milk protein polymorphisms in cattle: effect on animal breeding and human nutrition. J Dairy Sci 92, 5335-5352.

Castro, L.F.C., Gonçalves, O., Mazan, S., Tay, B.-H., Venkatesh, B., Wilson, J.M., 2014. Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history. Proceedings of the Royal Society B: Biological Sciences 281.

Castro, L.F.C., Lopes-Marques, M., Gonçalves, O., Wilson, J.M., 2012. The Evolution of Pepsinogen C Genes in Vertebrates: Duplication, Loss and Functional Diversification. PLoS ONE 7, e32852.

Coe, C.L., Kemnitz, J.W., Schneider, M.L., 1993. Vulnerability of placental antibody transfer and fetal complement synthesis to disturbance of the pregnant monkey. Journal of Medical Primatology 22, 294-300.

Coe, C.L., Lubach, G.R., 2014. Vital and vulnerable functions of the primate placenta critical for infant health and brain development. Frontiers in Neuroendocrinology 35, 439-446.

Coe, C.L., Lubach, G.R., Izard, K.M., 1994. Progressive improvement in the transfer of maternal antibody across the order Primates. American Journal of Primatology 32, 51-55.

Cruywagen, C.W., 1990. Effect of Curd Forming of Colostrum on Absorption of Immunoglobulin G in Newborn Calves. Journal of Dairy Science 73, 3287-3290.

Dallas, D.C., Murray, N.M., Gan, J., 2015. Proteolytic Systems in Milk: Perspectives on the Evolutionary Function within the Mammary Gland and the Infant. J Mammary Gland Biol Neoplasia 20, 133-147.

Foltmann, B., 1992. Chymosin: A short review on foetal and neonatal gastric proteases. Scandinavian Journal of Clinical and Laboratory Investigation 52, 65-79.

Foltmann, B., Harlow, K., Houen, G., Nielsen, P.K., Sangild, P., 1995. Comparative Investigations on Pig Gastric Proteases and Their Zymogens. In: Takahashi, K. (Ed.), Aspartic Proteinases: Structure, Function, Biology, and Biomedical Implications. Springer US, Boston, MA, pp. 41-51.

Foltmann, B., Jensen, A.L., Lønblad, P., Smidt, E., Axelsen, N.H., 1981. A developmental analysis of the production of chymosin and pepsin in pigs. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry 68, 9-13.

Furukawa, S., Kuroda, Y., Sugiyama, A., 2014. A Comparison of the Histological Structure of the Placenta in Experimental Animals. Journal of Toxicologic Pathology 27, 11-18.

Go, Y., Satta, Y., Takenaka, O., Takahata, N., 2005. Lineage-Specific Loss of Function of Bitter Taste Receptor Genes in Humans and Nonhuman Primates. Genetics 170, 313-326.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology 59, 307-321.

Hayashi, K., Agata, K., Mochii, M., Yasugi, S., Eguchi, G., Mizuno, T., 1988. Molecular Cloning and the Nucleotide Sequence of cDNA for Embryonic Chicken Pepsinogen: Phylogenetic Relationship with Prochymosin. Journal of Biochemistry 103, 290-296.

Hedges, S.B., Dudley, J., Kumar, S., 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22, 2971-2972.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., Kumar, S., 2015. Tree of Life Reveals Clock-Like Speciation and Diversification. Molecular Biology and Evolution 32, 835-845.

Hurley, W.L., Theil, P.K., 2011. Perspectives on Immunoglobulins in Colostrum and Milk. Nutrients 3, 442-474.

Jensen, T., Axelsen, N.H., Foltmann, B., 1982. Isolation and partial characterization of prochymosin and chymosin from cat. Biochimica et Biophysica Acta 705, 249-256.

Kageyama, T., 2000. New World Monkey Pepsinogens A and C, and Prochymosins. Purification, Characterization of Enzymatic Properties, cDNA Cloning, and Molecular Evolution. Journal of Biochemistry 127, 761-770.

Kageyama, T., 2002. Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. Cellular and Molecular Life Sciences CMLS 59, 288-306.

Kageyama, T., Ichinose, M., Tsukada-Kato, S., Omata, M., Narita, Y., Moriyama, A., Yonezawa, S., 2000. Molecular Cloning of Neonate/Infant-Specific Pepsinogens from Rat Stomach Mucosa and Their Expressional Change during Development. Biochemical and Biophysical Research Communications 267, 806-812.

Kageyama, T., Takahashi, K., 1984. Rabbit pepsinogens. European Journal of Biochemistry 141, 261-269.

Kageyama, T., Tanabe, K., Koiwai, O., 1990. Structure and development of rabbit pepsinogens. Stage-specific zymogens, nucleotide sequences of cDNAs, molecular evolution, and gene expression during development. Journal of Biological Chemistry 265, 17031-17038.

Kageyama, T., Tanabe, K., Koiwai, O., 1991. Development-dependent expression of isozymogens of monkey pepsinogens and structural differences between them. European Journal of Biochemistry 202, 205-215.

Katoh, K., Kuma, K.-i., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 33, 511-518.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings in bioinformatics 9, 286-298.

Kawasaki, K., Weiss, K.M., 2003. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. Proc Natl Acad Sci U S A 100, 4060-4065.

Kohl, S., Loo, L.S., 1984. The Relative Role of Transplacental and Milk Immune Transfer in Protection Against Lethal Neonatal Herpes Simplex Virus Infection in Mice. Journal of Infectious Diseases 149, 38-42.

Kolmer, M., Örd, T., Alhonen, L., Hyttinen, J.-M., Saarma, M., Villems, R., Jänne, J., 1991. Assignment of human prochymosin pseudogene to chromosome 1. Genomics 10, 496-498.

Kumar, S., Hedges, S.B., 2011. TimeTree2: species divergence times on the iPhone. Bioinformatics 27, 2023-2024.

Langer, P., 2009. Differences in the Composition of Colostrum and Milk in Eutherians Reflect Differences in Immunoglobulin Transfer. Journal of Mammalogy 90, 332-339.

Langholm Jensen, J., Mølgaard, A., Navarro Poulsen, J.-C., Harboe, M.K., Simonsen, J.B., Lorentzen, A.M., Hjernø, K., van den Brink, J.M., Qvist, K.B., Larsen, S., 2013. Camel and bovine chymosin: the relationship between their structures and cheese-making properties. Acta Crystallographica Section D: Biological Crystallography 69, 901-913.

Mercier, J.-C., Chobert, J.-M., Addeo, F., 1976. Comparative study of the amino acid sequences of the caseinomacropeptides from seven species. FEBS Letters 72, 208-214.

Mokhber-Dezfooli, M.R., Nouri, M., Rasekh, M., Constable, P.D., Effect of abomasal emptying rate on the apparent efficiency of colostral immunoglobulin G absorption in neonatal Holstein-Friesian calves. Journal of Dairy Science 95, 6740-6749.

Narita, Y., Oda, S.-i., Takenaka, O., Kageyama, T., 2010. Lineage-Specific Duplication and Loss of Pepsinogen Genes in Hominoid Evolution. Journal of Molecular Evolution 70, 313-324.

Nei, M., Glazko, G.V., 2002. The Wilhelmine E. Key 2001 Invitational Lecture. Estimation of Divergence Times for a Few Mammalian and Several Primate Species. Journal of Heredity 93, 157-164.

Ord T, K.M., Villems R, Saarma M., 1990. Structure of the human genomic region homologous to the bovine prochymosin-encoding gene. 91, 6.

Örd, T., Kolmer, M., Villems, R., Saarma, M., 1990. Structure of the human genomic region homologous to the bovine prochymosin-encoding gene. Gene 91, 241-246.

Ordoñez, G.R., Hillier, L.W., Warren, W.C., Grützner, F., López-Otín, C., Puente, X.S., 2008. Loss of genes implicated in gastric function during platypus evolution. Genome Biology 9, R81-R81.

Palmer, D.S., Christensen, A.U., Sørensen, J., Celik, L., Qvist, K.B., Schiøtt, B., 2010. Bovine Chymosin: A Computational Study of Recognition and Binding of Bovine κ-Casein. Biochemistry 49, 2563-2573.

Pearl, L., Blundell, T., 1984. The active site of aspartic proteinases. FEBS Letters 174, 96-101.

Pentšuk, N., van der Laan, J.W., 2009. An interspecies comparison of placental antibody transfer: New insights into developmental toxicity testing of monoclonal antibodies. Birth Defects Research Part B: Developmental and Reproductive Toxicology 86, 328-344.

Pond, S.L.K., Frost, S.D.W., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21, 2531-2533.

Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676-679.

Pungerčar, J., Štrukelj, B., Gubenšek, F., Turk, V., Kregar, I., 1991. Amino Acid Sequence of Lamb Preprochymosin and its Comparison to Other Chymosins. In: Dunn, B.M. (Ed.), Structure and Function of the Aspartic Proteinases: Genetics, Structures, and Mechanisms. Springer US, Boston, MA, pp. 127-131.

Rojas, R., Apodaca, G., 2002. Immunoglobulin transport across polarized epithelial cells. Nat Rev Mol Cell Biol 3, 944-955.

Schrago, C.G., Russo, C.A.M., 2003. Timing the Origin of New World Monkeys. Molecular Biology and Evolution 20, 1620-1625.

Steiner, C.C., Ryder, O.A., 2011. Molecular phylogeny and evolution of the Perissodactyla. Zoological Journal of the Linnean Society 163, 1289-1303.

Szecsi, P.B., 1992. The aspartic proteases. Scandinavian Journal of Clinical and Laboratory Investigation 52, 5-22.

Uniacke-Lowe, T., Huppertz, T., Fox, P.F., 2010. Equine milk proteins: Chemistry, structure and nutritional significance. International Dairy Journal 20, 609-629.

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., Scheffler, K., 2014. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution.

Wu, C.I., Li, W.H., Shen, J.J., Scarpulla, R.C., Limbach, K.J., Wu, R., 1986. Evolution of cytochrome c genes and pseudogenes. Journal of Molecular Evolution 23, 61-75.

Wu, T., Sun, L.-C., Du, C.-H., Cai, Q.-F., Zhang, Q.-B., Su, W.-J., Cao, M.-J., 2009. Identification of pepsinogens and pepsins from the stomach of European eel (*Anguilla anguilla*). Food Chemistry 115, 137-142.

Yakabe, E., Tanji, M., Ichinose, M., Goto, S., Miki, K., Kurokawa, K., Ito, H., Kageyama, T., Takahashi, K., 1991. Purification, characterization, and amino acid sequences of pepsinogens and pepsins from the esophageal mucosa of bullfrog (*Rana catesbeiana*). Journal of Biological Chemistry 266, 22436-22443.

## Supplementary Material

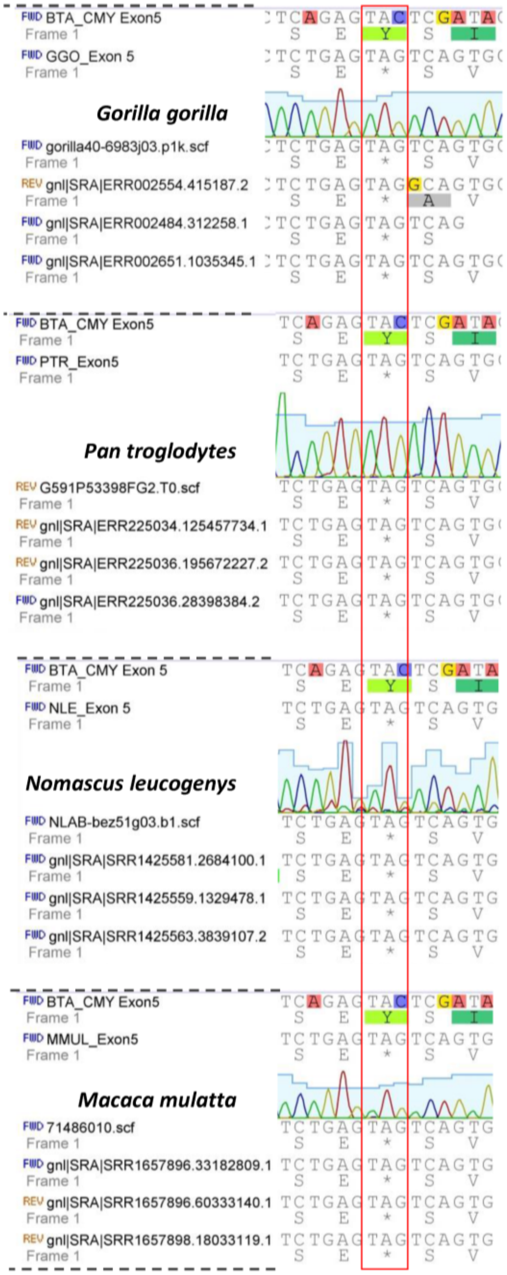
**Table 1:** Accession numbers of the analysed sequences * indicates candidate pseudogenes

| | | Species | Order | Accession number |
|---|---|---|---|---|
| 1 | HSA | *Homo sapiens* | Primate-Hominoidae | NR_003599 * |
| 2 | GGO | *Gorilla gorilla gorilla* | Primate-Hominoidae | XM_004026309.1 * |
| 3 | PTR | *Pan troglodytes* | Primate-Hominoidae | LOC748852 * |
| 4 | PAB | *Pongo abelii* | Primate-Hominoidae | LOC100445768 * |
| 5 | PPA | *Pan paniscus* | Primate-Hominoidae | LOC100971639* |
| 6 | NLE | *Nomascus leucogenys* | Primate-Hominoidae | XM_003267945.1* |
| 7 | MLE | *Mandrillus leucophaeus* | Primate-Cercopithecoidea | XM_011969946.1 * |
| 8 | MNE | *Macaca nemestrina* | Primate-Cercopithecoidea | XM_011737158.1 * |
| 9 | MMUL | *Macaca mulatta* | Primate-Cercopithecoidea | XM_015147525.1 * |
| 10 | MFA | *Macaca fascicularis* | Primate-Cercopithecoidea | XM_005545183.2 * |
| 11 | RRO | *Rhinopithecus roxellana* | Primate-Cercopithecoidea | XM_010386267.1* |
| 12 | RBI | *Rhinopithecus bieti* | Primate-Cercopithecoidea | XM_017869581.1 * |
| 13 | CAT | *Cercocebus atys* | Primate-Cercopithecoidea | XM_012078338.1* |
| 14 | CSA | *Chlorocebus sabaeus* | Primate-Cercopithecoidea | XM_007977631.1* |
| 15 | CAN | *Colobus angolensis palliatus* | Primate-Cercopithecoidea | XM_011948155.1* |
| 16 | PAN | *Papio anubis* | Primate-Cercopithecoidea | LOC101006467* |
| 17 | ANA | *Aotus nancymaae* | Primate-Platyrrhini | XM_012468016.1 * |
| 18 | OGA | *Otolemur garnettii* | Primate-Platyrrhini | XM_003784040 |
| 19 | CCA | *Cebus capucinus* | Primate-Platyrrhini | XM_017499374.1 |
| 20 | SBO | *Saimiri boliviensis* | Primate | XM_003933436.2 |
| 21 | CJA | *Callithrix jacchus* | Primate | XM_002751197.2 |
| 22 | TSY | *Tarsius syrichta* | Primate | XM_008068475.1 |
| 23 | PCO | *Propithecus coquereli* | Primate | XM_012663376.1 |
| 24 | MMUR | *Microcebus murinus* | Primate | XM_012751547.1 |
| 25 | NGA | *Nannospalax galili* | Rodentia-Myomorpha | XM_008850353 |
| 26 | CGR | *Cricetulus griseus* | Rodentia-Myomorpha | XM_007621724.1 |
| 27 | PMA | *Peromyscus maniculatus bairdii* | Rodentia-Myomorpha | XM_006979512.2 |
| 28 | MOC | *Microtus ochrogaster* | Rodentia-Myomorpha | XM_005367834.1 |
| 29 | MMU | *Mus musculus* | Rodentia-Myomorpha | NM_001111143.1 |
| 30 | RNO | *Rattus norvegicus* | Rodentia-Myomorpha | NM_020091.1 |
| 31 | JJA | *Jaculus jaculus* | Rodentia-Myomorpha | XM_004659006.1 |
| 32 | ITR | *Ictidomys tridecemlineatus* | Rodentia-Sciuromorpha | XM_005338138.1 |
| 33 | DOR | *Dipodomys ordii* | Rodentia-Castorimorpha | XM_013031883_1 |
| 34 | MMA | *Marmota marmota marmota* | Rodentia-Sciuromorpha | XM_015505064.1 * |
| 35 | CPO | *Cavia porcellus* | Rodentia-Hystricomorpha | Not annotated * |
| 36 | MAU | *Mesocricetus auratus* | Rodentia-Myomorpha | XM_013118081.1* |
| 37 | HGL | *Heterocephalus glaber* | Rodentia-Hystricomorpha | Not annotated * |
| 38 | ODE | *Octodon degus* | Rodentia-Hystricomorpha | Not annotated * |
| 39 | FDA | *Fukomys damarensis* | Rodentia-Hystricomorpha | Not annotated * |
| 40 | CLA | *Chinchilla lanigera* | Rodentia-Hystricomorpha | Not annotated * |
| 41 | OCU | *Oryctolagus cuniculus* | Lagomorpha | XM_002715787.2 |
| 42 | OPR | *Ochotona princeps* | Lagomorpha | XM_004581942.1 |
| 43 | SSC | *Sus scrofa* | Artiodactyla | XM_001927061.6 |
| 44 | VPA | *Vicugna pacos* | Artiodactyla | XM_006197438.1 |
| 45 | CFE | *Camelus ferus* | Artiodactyla | XM_006194757.1 |
| 46 | CBA | *Camelus bactrianus* | Artiodactyla | XM_010948280.1 |
| 47 | CDR | *Camelus dromedarius* | Artiodactyla | AJ131677.1 |
| 48 | BMU | *Bos mutus* | Artiodactyla | XM_005891276.2 |
| 49 | BBI | *Bison bison bison* | Artiodactyla | XM_010830872.1 |
| 50 | BGR | *Bos grunniens* | Artiodactyla | JX839990.1 |
| 51 | BTA | *Bos taurus* | Artiodactyla | NM_180994.2 |
| 52 | BBU | *Bubalus bubalis* | Artiodactyla | XM_006064953.1 |
| 53 | PHO | *Pantholops hodgsonii* | Artiodactyla | XM_005985276.1 |
| 54 | OAR | *Ovis aries* | Artiodactyla | NM_001009804.1 |
| 55 | CHI | *Capra hircus* | Artiodactyla | NM_001285759.1 |
| 56 | PCA | *Physeter catodon* | Cetacea | XM_007106490.1 |
| 57 | LVE | *Lipotes vexillifer* | Cetacea | XM_007449289.1 |

| | | | | |
|---|---|---|---|---|
| 58 | BAC | *Balaenoptera acutorostrata scammoni* | Cetacea | XM_007169436.1 |
| 59 | OOR | *Orcinus orca* | Cetacea | XM_004263168.2* |
| 60 | TMA | *Trichechus manatus latirostris* | Sirenia | XM_004390419.1 |
| 61 | CSI | *Ceratotherium simum simum* | Perissodactyla-Rhinoceratidae | LOC101406254* |
| 62 | ECA | *Equus caballus* | Perissodactyla-Equidea | XM_014740013.1* |
| 63 | EPR | *Equus przewalskii* | Perissodactyla-Equidea | LOC103556180* |
| 64 | ORO | *Odobenus rosmarus divergens* | Carnivora | XM_004407156.1 |
| 65 | LWE | *Leptonychotes weddellii* | Carnivora | XM_006749503.1* |
| 66 | AME | *Ailuropoda melanoleuca* | Carnivora | XM_002928298.2 |
| 67 | UMA | *Ursus maritimus* | Carnivora | XM_008704042.1 * |
| 68 | MFU | *Mustela putorius furo* | Carnivora | XM_004769748.1 |
| 69 | CFA | *Canis lupus familiaris* | Carnivora | XM_003639133.1 |
| 70 | PALT | *Panthera tigris altaica* | Carnivora | XM_007076400.2 |
| 71 | FCA | *Felis catus* | Carnivora | XM_003990428.1 |
| 72 | AJU | *Acinonyx jubatus* | Carnivora | XM_015075832.1 |
| 73 | PVA | *Pteropus vampyrus* | Chiroptera | XM_011369072.1 |
| 74 | EFU | *Eptesicus fuscus* | Chiroptera | XM_008147352.1 |
| 75 | PALE | *Pteropus alecto* | Chiroptera | XM_006919623.1 |
| 76 | MDA | *Myotis davidii* | Chiroptera | XM_006762989.2 |
| 77 | MBR | *Myotis brandtii* | Chiroptera | XM_005860042.2 |
| 78 | MNA | *Miniopterus natalensis* | Chiroptera | XM_016223323.1 |
| 79 | MLU | *Myotis lucifugus* | Chiroptera | XM_014458936.1 |
| 80 | ETE | *Echinops telfairi* | Tenrecidae | XM_004699199.1 |
| 81 | CAS | *Chrysochloris asiatica* | Afrosoricida | XM_006865933.1 |
| 82 | OAF | *Orycteropus afer afer* | Tubulidentata | XM_007949443.1 |
| 83 | LAF | *Loxodonta africana* | Proboscidea | XM_003409400.2 |
| 84 | EED | *Elephantulus edwardii* | Macroscelidea | XM_006899299.1 |
| 85 | TCH | *Tupaia chinensis* | Scandentia | XM_006149562.1 |
| 86 | GVA | *Galeopterus variegatus* | Dermoptera | XM_008573253* |
| 87 | MJA | *Manis javanica* | Pholidota | XM_017664973.1 |
| 88 | CCR | *Condylura cristata* | Soricomorpha | XM_004689665.1 |
| 89 | SAR | *Sorex araneus* | Eulipotyphla | XM_004620085.1 |
| 90 | EEU | *Erinaceus europaeus* | Erinaceomorpha | XM_007529363.1 |
| 91 | DNO | *Dasypus novemcinctus* | Cingulata | XM_004484022.1 |
| 92 | SHA | *Sarcophilus harrisii* | Marsupialia | XM_003769757.1 |
| 93 | MDO | *Monodelphis domestica* | Marsupialia | XM_001372646.3 |
| 94 | OAN | *Ornithorhynchus anatinus* | Monotremata | XM_016228296.1* |
| 95 | GGA | *Gallus gallus* | Birds | ENSGALT00000000594.1 |
| 96 | MGA | *Meleagris gallopavo* | Birds | ENSMGAT00000002636.2 |
| 97 | APL | *Anas platyrhynchos* | Birds | ENSAPLT00000016539.1 |
| 98 | FAL | *Ficedula albicollis* | Birds | ENSFALT00000003260.1 |
| 99 | ACA | *Anolis carolinensis* | Reptile | ENSACAT00000013099.3 |

## Supplementary material 1

SRA and Trace archives for stop mutation in exon 5 in Hominoidea and Cercopithecoidea


FWD BTA_CMY Exon5
Frame 1
FWD GGO_Exon 5
Frame 1
Gorilla gorilla
FWD gorilla40-6983j03.p1k.scf
REV gnl|SRA|ERR002554.415187.2
FWD gnl|SRA|ERR002484.312258.1
FWD gnl|SRA|ERR002651.1035345.1
FWD PTR_Exon5
Pan troglodytes
REV G591P53398FG2.T0.scf
REV gnl|SRA|ERR225034.125457734.1
REV gnl|SRA|ERR225036.195672227.2
FWD gnl|SRA|ERR225036.28398384.2
FWD BTA_CMY Exon 5
FWD NLE_Exon 5
Nomascus leucogenys
FWD NLAB-bez51g03.b1.scf
FWD gnl|SRA|SRR1425581.2684100.1
FWD gnl|SRA|SRR1425559.1329478.1
FWD gnl|SRA|SRR1425563.3839107.2
FWD MMUL_Exon5
Macaca mulatta
FWD 71486010.scf
FWD gnl|SRA|SRR1657896.33182809.1
REV gnl|SRA|SRR1657896.60333140.1
REV gnl|SRA|SRR1657898.18033119.1
TCAGAGTACTCGATA
S E Y S I
TCTGAGTAGTCAGTG
S E * S V

Run ERR002554- ERX000192 release date: 2010-03-15
Run ERR002484- ERX000198 release date: 2010-03-15
Run ERR002651- ERX000200 release date. 2010-03-15

Run ERR225034- ERX199694 release date: 2013-04-08
Run ERR225036- ERX199696 release date: 2013-04-08

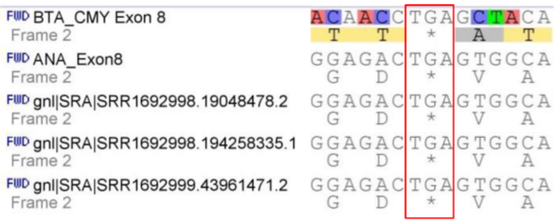
Run SRR1425581- SRX590192 release date: 2014-09-11
Run SRR1425559- SRX590192 release date: 2014-09-11
Run SRR1425563- SRX590192 release date: 2014-09-11

Run SRR1425581- SRX590192 release date: 2014-09-11
Run SRR1425559- SRX590192 release date: 2014-09-11
Run SRR1425563- SRX590192 release date: 2014-09-11

## Supplementary material 2

SRA for stop mutation in exon 8 in *Aotus nancymaae*

***Aotus nancymaae***

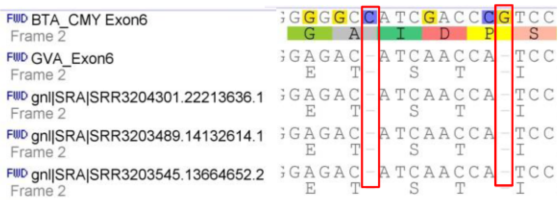


```
FWD BTA_CMY Exon 8                    ACAACCTGAGCTACA
Frame 2                                T  T  *  A  T
FWD ANA_Exon8                         GGAGACTGAGTGGCA
Frame 2                                G  D  *  V  A
FWD gnl|SRA|SRR1692998.19048478.2     GGAGACTGAGTGGCA
Frame 2                                G  D  *  V  A
FWD gnl|SRA|SRR1692998.194258335.1    GGAGACTGAGTGGCA
Frame 2                                G  D  *  V  A
FWD gnl|SRA|SRR1692999.43961471.2     GGAGACTGAGTGGCA
Frame 2                                G  D  *  V  A
```

Run SRR1692998- SRX795827 release date: 2014-12-05
Run SRR1692999- SRX795827 release date: 2014-12-05

## Supplementary material 3

SRA for *Galeopterus variegatus* 2nt deletion in exon 5



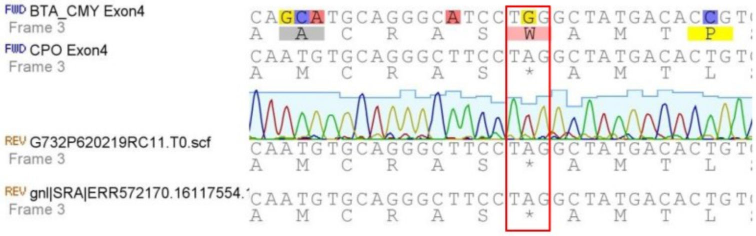
```
FWD BTA_CMY Exon6                     GGGGCCATCGACCCGTCC
Frame 2                                G  A  I  D  P  S
FWD GVA_Exon6                         GGAGAC-ATCAACCA-TCC
Frame 2                                E  T-   S  T  -I
FWD gnl|SRA|SRR3204301.22213636.1     GGAGAC-ATCAACCA-TCC
Frame 2                                E  T-   S  T  -I
FWD gnl|SRA|SRR3203489.14132614.1     GGAGAC-ATCAACCA-TCC
Frame 2                                E  T-   S  T  -I
FWD gnl|SRA|SRR3203545.13664652.2     GGAGAC-ATCAACCA-TCC
Frame 2                                E  T    S  T  -I
```

Run SRR3204301- SRX1612886 release date: 2016-03-04
Run SRR3203489- SRX1612894 release date: 2016-03-04

## Supplementary material 4

SRA for Rodentia

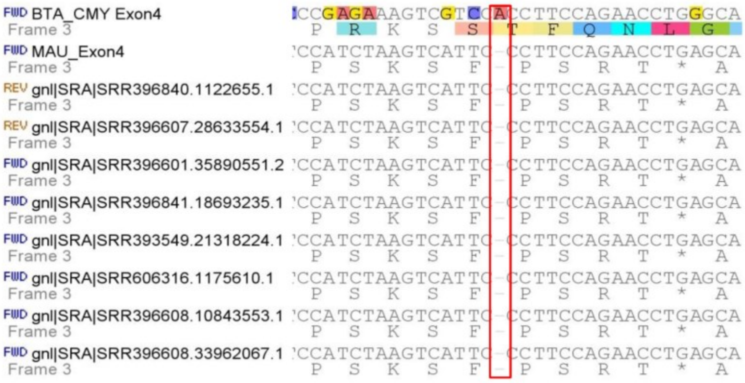
### *Cavia porcellus*



```
FWD BTA_CMY Exon4                 CAGCATGCAGGGCATCCTGGGCTATGACACCGT
Frame 3                           A  A  C  R  A  S  W  A  M  T  P
FWD CPO Exon4                     CAATGTGCAGGGCTTCCTAGGCTATGACACTGT
Frame 3                           A  M  C  R  A  S  *  A  M  T  L
REV G732P620219RC11.T0.scf        CAATGTGCAGGGCTTCCTAGGCTATGACACTGT
Frame 3                           A  M  C  R  A  S  *  A  M  T  L
REV gnl|SRA|ERR572170.16117554.   CAATGTGCAGGGCTTCCTAGGCTATGACACTGT
Frame 3                           A  M  C  R  A  S  *  A  M  T  L
```

Run ERR572170- ERX530898 release date: 2015-02-03

### *Mesocricetus auratus*



```
FWD BTA_CMY Exon4                 CCGAGAAAGTCGTCCACCTTCCAGAACCTGGGCA
Frame 3                            P  R  K  S  S  T  F  Q  N  L  G
FWD MAU_Exon4                     CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
REV gnl|SRA|SRR396840.1122655.1   CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
REV gnl|SRA|SRR396607.28633554.1  CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
FWD gnl|SRA|SRR396601.35890551.2  CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
FWD gnl|SRA|SRR396841.18693235.1  CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
FWD gnl|SRA|SRR393549.21318224.1  CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
FWD gnl|SRA|SRR606316.1175610.1   CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
FWD gnl|SRA|SRR396608.10843553.1  CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
FWD gnl|SRA|SRR396608.33962067.1  CCATCTAAGTCATTC-CCTTCCAGAACCTGAGCA
Frame 3                            P  S  K  S  F  -  P  S  R  T  *  A
```

Run SRR396840/SRR396601-SRX114570 release date: 2012-01-07

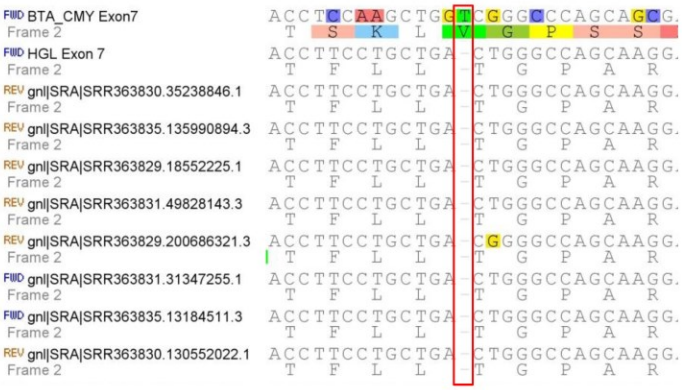
Run SRR396607/SRR396841-SRX114568 release date: 2012-01-05

Run SRR393549- SRX113462 release date: 2011-12-29

Run SRR606316- SRX200267 release date: 2012-10-23

Run SRR396608- SRX114572 release date: 2012-01-05

### *Heterocephalus glaber*

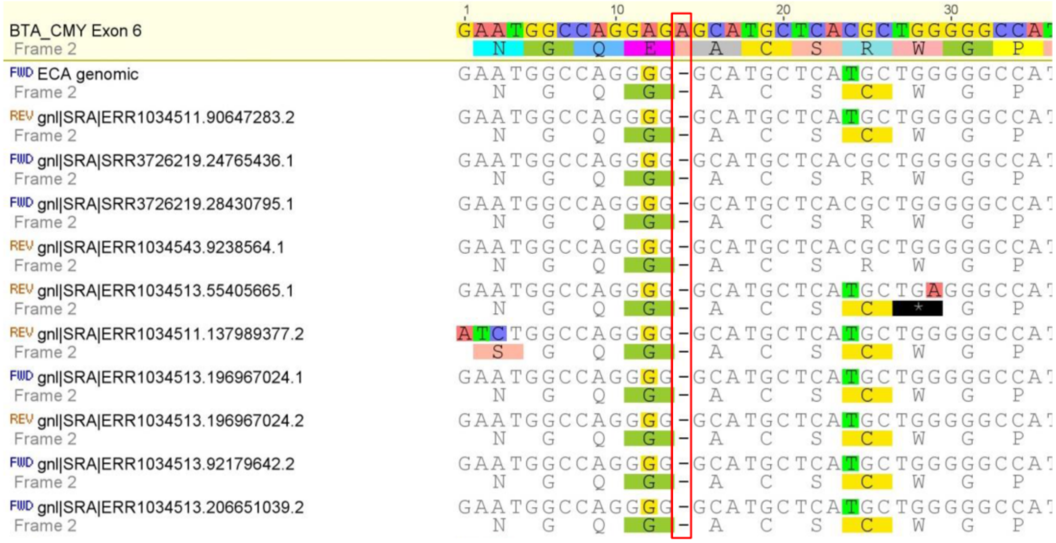


```
FWD BTA_CMY Exon7                 ACCTCCAAGCTGGTCGGGCCAGCAGCG.
Frame 2                             T  S  K  L  V  G  P  S  S
FWD HGL Exon 7                    ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
REV gnl|SRA|SRR363830.35238846.1  ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
REV gnl|SRA|SRR363835.135990894.3 ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
REV gnl|SRA|SRR363829.18552225.1  ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
REV gnl|SRA|SRR363831.49828143.3  ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
REV gnl|SRA|SRR363829.200686321.3 ACCTTCCTGCTGA-CGGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
FWD gnl|SRA|SRR363831.31347255.1  ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
FWD gnl|SRA|SRR363835.13184511.3  ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
REV gnl|SRA|SRR363830.130552022.1 ACCTTCCTGCTGA-CTGGGCCAGCAAGG.
Frame 2                             T  F  L  L  -T  G  P  A  R
```

Run SRR363830/SRR363835/SRR363829/SRR363831-SRX105046
release date: 2011-11-10

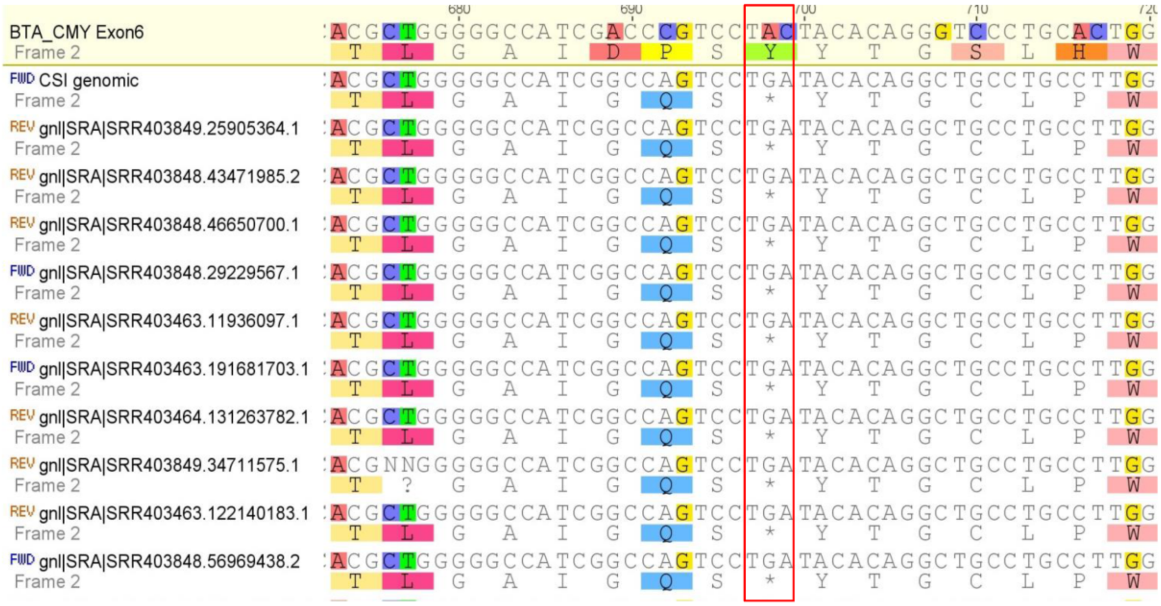
## Supplementary material 5

SRA for *Equus caballus* 1nt deletion in exon 6



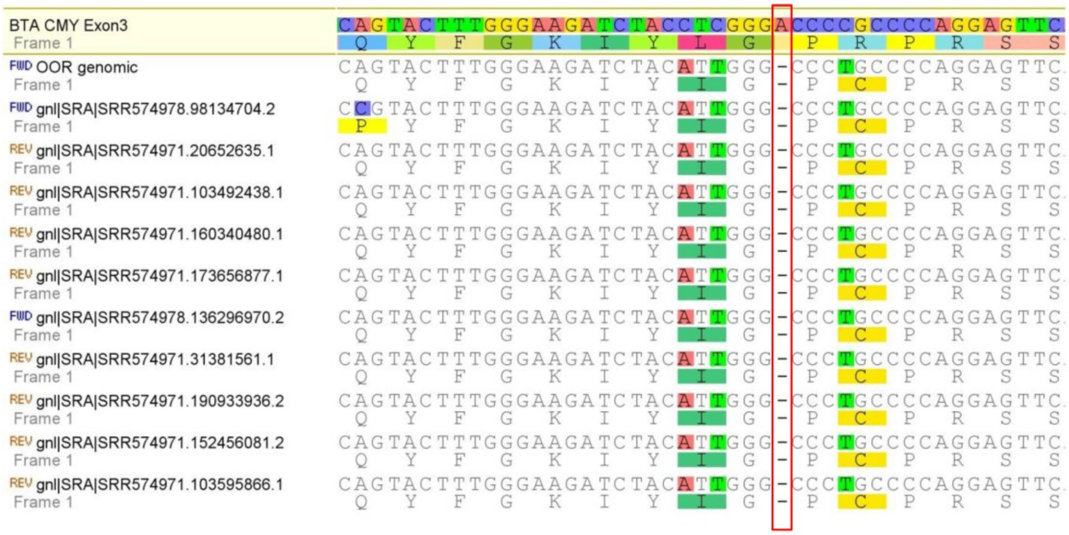
```
                                    1        10        20        30
BTA_CMY Exon 6                      GAATGGCCAGGAGAGCATGCTCACGCTGGGGGCCAT
 Frame 2                             N  G  Q  E   A  C  S  R  W  G  P
FWD ECA genomic                     GAATGGCCAGGGG-GCATGCTCATGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  C  W  G  P
REV gnl|SRA|ERR1034511.90647283.2   GAATGGCCAGGGG-GCATGCTCATGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  C  W  G  P
FWD gnl|SRA|SRR3726219.24765436.1   GAATGGCCAGGGG-GCATGCTCACGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  R  W  G  P
FWD gnl|SRA|SRR3726219.28430795.1   GAATGGCCAGGGG-GCATGCTCACGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  R  W  G  P
REV gnl|SRA|ERR1034543.9238564.1    GAATGGCCAGGGG-GCATGCTCACGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  R  W  G  P
REV gnl|SRA|ERR1034513.55405665.1   GAATGGCCAGGGG-GCATGCTCATGCTGAGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  C  *  G  P
REV gnl|SRA|ERR1034511.137989377.2  ATCTGGCCAGGGG-GCATGCTCATGCTGGGGGCCAT
 Frame 2                             S  G  Q  G - A  C  S  C  W  G  P
FWD gnl|SRA|ERR1034513.196967024.1  GAATGGCCAGGGG-GCATGCTCATGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  C  W  G  P
REV gnl|SRA|ERR1034513.196967024.2  GAATGGCCAGGGG-GCATGCTCATGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  C  W  G  P
FWD gnl|SRA|ERR1034513.92179642.2   GAATGGCCAGGGG-GCATGCTCATGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  C  W  G  P
FWD gnl|SRA|ERR1034513.206651039.2  GAATGGCCAGGGG-GCATGCTCATGCTGGGGGCCAT
 Frame 2                             N  G  Q  G - A  C  S  C  W  G  P
```

SRA for *Ceratotherium simum* stop codon in exon 6



```
                                         680       690       700       710       720
BTA_CMY Exon6                        ACGCTGGGGGCCATCGACCCGTCCTACTACACAGGTCCCTGCACTGG
 Frame 2                              T  L  G  A  I  D  P  S  Y  Y  T  G  S  L  H  W
FWD CSI genomic                      ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
REV gnl|SRA|SRR403849.25905364.1     ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
REV gnl|SRA|SRR403848.43471985.2     ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
REV gnl|SRA|SRR403848.46650700.1     ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
FWD gnl|SRA|SRR403848.29229567.1     ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
REV gnl|SRA|SRR403463.11936097.1     ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
FWD gnl|SRA|SRR403463.191681703.1    ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
REV gnl|SRA|SRR403464.131263782.1    ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
REV gnl|SRA|SRR403849.34711575.1     ACGNNGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  ?  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
REV gnl|SRA|SRR403463.122140183.1    ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
FWD gnl|SRA|SRR403848.56969438.2     ACGCTGGGGGCCATCGGCCAGTCCTGATACACAGGCTGCCTGCCTTGG
 Frame 2                              T  L  G  A  I  G  Q  S  *  Y  T  G  C  L  P  W
```

**Supplementary material 6**
SRA for *Orcinus Orca* 1nt deletion exon 3

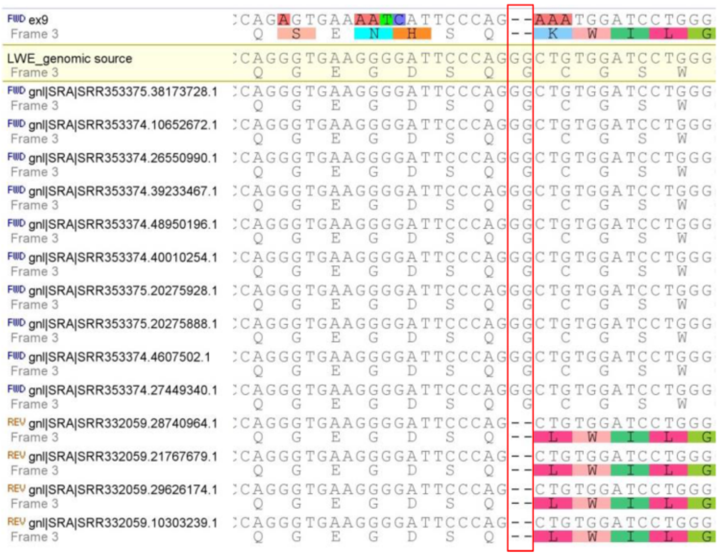


```
BTA CMY Exon3                        CAGTACTTTGGGAAGATCTACCTCGGGACCCCGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  L  G   P  R  P  R  S  S
FWD OOR genomic                      CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
FWD gnl|SRA|SRR574978.98134704.2     CCGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             P  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.20652635.1     CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.103492438.1    CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.160340480.1    CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.173656877.1    CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
FWD gnl|SRA|SRR574978.136296970.2    CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.31381561.1     CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.190933936.2    CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.152456081.2    CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
REV gnl|SRA|SRR574971.103595866.1    CAGTACTTTGGGAAGATCTACATTGGG-CCCTGCCCCAGGAGTTC
  Frame 1                             Q  Y  F  G  K  I  Y  I  G - P  C  P  R  S  S
```

**Supplementary material 7**

SRA for *Leptonychotes weddellii* (LWE) and *Ursus maritimus* (UMA)

### *Leptonychotes weddellii* 2bp frameshift exon9

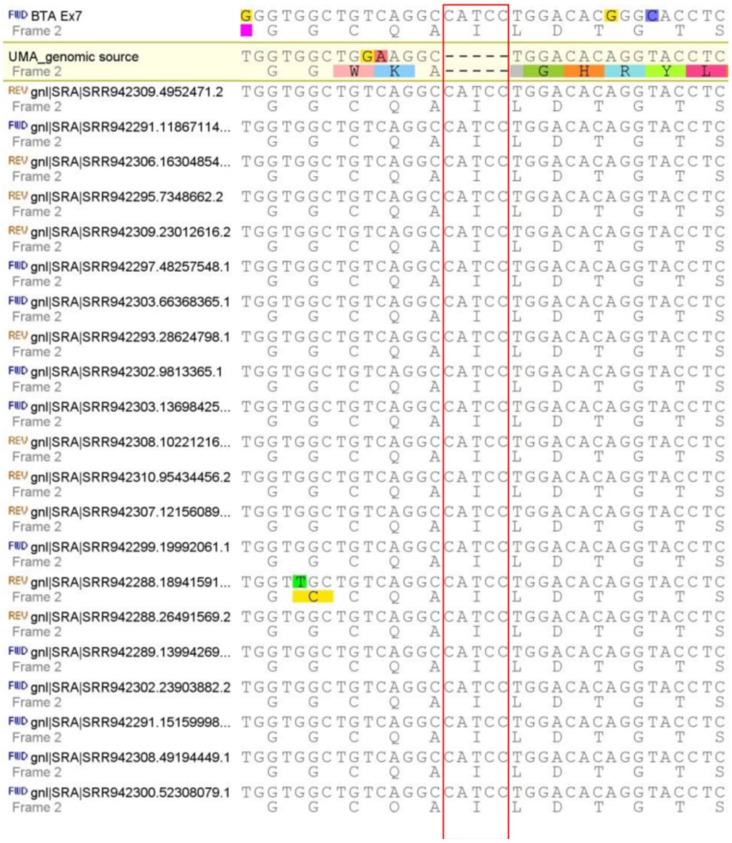


```
FWD ex9                          CAGAGTGAAAATCATTCCCAG--AAATGGATCCTGGG
 Frame 3                          Q  S  E  N  H  S  Q -- K  W  I  L  G
LWE_genomic source               CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353375.38173728.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353374.10652672.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353374.26550990.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353374.39233467.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353374.48950196.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353374.40010254.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353375.20275928.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353375.20275888.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353374.4607502.1  CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
FWD gnl|SRA|SRR353374.27449340.1 CAGGGTGAAGGGGATTCCCAGGGCTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q  G  C  G  S  W
REV gnl|SRA|SRR332059.28740964.1 CAGGGTGAAGGGGATTCCCAG--CTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q -- L  W  I  L  G
REV gnl|SRA|SRR332059.21767679.1 CAGGGTGAAGGGGATTCCCAG--CTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q -- L  W  I  L  G
REV gnl|SRA|SRR332059.29626174.1 CAGGGTGAAGGGGATTCCCAG--CTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q -- L  W  I  L  G
REV gnl|SRA|SRR332059.10303239.1 CAGGGTGAAGGGGATTCCCAG--CTGTGGATCCTGGG
 Frame 3                          Q  G  E  G  D  S  Q -- L  W  I  L  G
```

Run SRR332059- SRX091973 release date: 2011-08-22

Run SRR353374- SRX101110 release date: 2011-10-16

### *Ursus maritimus* 5 bp frameshift exon7



```
FWD BTA Ex7                         GGGTGGCTGTCAGGCCATCCTGGACACGGGCACCTC
 Frame 2                            ■ G  G  C  Q  A  I  L  D  T  G  T  S
UMA_genomic source                  TGGTGGCTGGAAGGC-----TGGACACAGGTACCTC
 Frame 2                             G  G  W  K  A-----  G  H  R  Y  L
REV gnl|SRA|SRR942309.4952471.2     TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942291.11867114...   TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942306.16304854...   TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942295.7348662.2     TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942309.23012616.2    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942297.48257548.1    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942303.66368365.1    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942293.28624798.1    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942302.9813365.1     TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942303.13698425...   TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942308.10221216...   TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942310.95434456.2    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942307.12156089...   TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942299.19992061.1    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942288.18941591...   TGGTTGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  C  C  Q  A  I  L  D  T  G  T  S
REV gnl|SRA|SRR942288.26491569.2    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942289.13994269...   TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942302.23903882.2    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942291.15159998...   TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942308.49194449.1    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
FWD gnl|SRA|SRR942300.52308079.1    TGGTGGCTGTCAGGCCATCCTGGACACAGGTACCTC
 Frame 2                             G  G  C  Q  A  I  L  D  T  G  T  S
```

Run SRR942309- SRX327155 release date: 2015-07-22

Run SRR942291- SRX327140 release date: 2015-07-22

Run SRR942306- SRX327153 release date: 2015-07-22

Run SRR942295- SRX327144 release date: 2015-07-22

Run SRR942297- SRX327146 release date: 2015-07-22

Run SRR942303- SRX327152 release date: 2015-07-22

Run SRR942293- SRX327142 release date: 2015-07-22

# CHAPTER VII

## DISCUSSION

# CHAPTER VII - DISCUSSION

Genome/gene duplication, loss, mutation, among other genetic factors play a fundamental role in animal evolution. Duplication has been referred as a major driver in vertebrate evolution and diversity (Ohno, 1970; Shimeld *et al.*, 2000; Cañestro, 2012; Chen *et al.*, 2013). As mentioned in the Chapter I, duplication leads to the creation of redundant genetic material, and following duplication purifying selection has been found to relax in at least one of the duplicates, allowing for the accumulation of mutations and consequent asymmetrical evolution of the duplicate genes (Ohno, 1970; Zhang, 2003). This asymmetrical evolution often ends with the loss of one duplicate, by the accumulation of deleterious mutations. However, occasionally one duplicate gene may accumulate mutations that are beneficial leading to its preservation and consequently to the retention of a larger genetic repertoire (Zhang, 2003; Louis, 2007).

This thesis has focused on the impact of these genomic events on the evolution of FA and protein metabolism in chordates, by the characterization of the corresponding genetic repertoires in the main chordate lineages. Yet, to fully perceive the evolution of FA metabolism and protein digestion in chordates it was essential to include information regarding the life history of the analyzed species such as: diet, trophic level, environment, transitions from aquatic to terrestrial or/and from marine to freshwater; which are all major players in the modulation of the genetic machinery involved in these pathways (Fig.1).

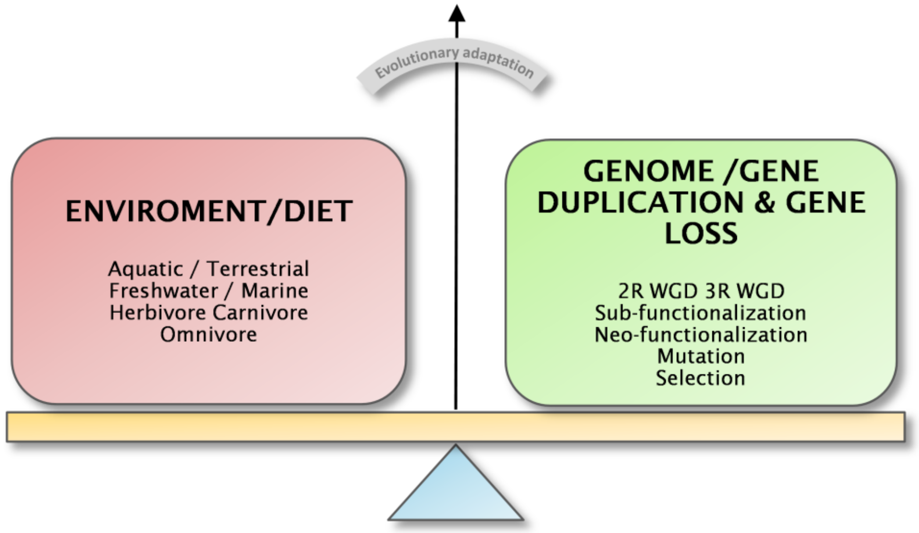


**Figure 1:** Schematic representation of the interplay between gene/genome duplication and gene loss with environmental factors, diet, all of which play a significant role in phenotypical outcome and evolutionary adaptation.

Here, taking advantage of the ever increasing genomic information available in public databases, the access to species placed at phylogenetic key positions, an integrative approach was pursued in the analysis of the of FA evolution in chordates. To this end, when possible, the genetic repertoire of the analyzed gene family and species was combined with functional characterization, and/or gene expression and/or information regarding environmental factors and dietary preferences.

### VII.1 FATTY ACID ACTIVATION

Before enrolling in several metabolic pathways such as FA biosynthesis, or β-oxidation, FAs require to be activated through the formation of thioesters with CoA (Watkins, 1997). This critical step is carried out by the Acyl-CoA synthetases enzymes; although these enzymes are well characterized regarding substrate specificity and expression, (Watkins, 1997; Watkins *et al.*, 2007) the evolutionary history and distribution of Acyl-CoA synthetases in vertebrates remained mostly unexplored.

The investigation of the Acyl-CoA short chain synthetase1 (*Acss1*) and the multigene family Acyl-CoA long chain synthetase (*Acsl*) reveals that the expansion of these gene families coincided with the 2R WGD and 3R WGD (Fig.2A (Chapter III)). This expansion was later followed by gene loss leading to the retention of distinct genetic repertoires in several lineages (Chapter III). Briefly, the analysis of *Acss1* previously assumed as a single member gene family in humans and mouse (Watkins *et al.*, 2007), uncovers a uncharacterized paralogous *Acss1* genes (named *Acss1b*) in several teleost species, birds and reptiles (Chapter III.1 (Castro *et al.*, 2012b)). Next, the analysis of the Acyl-CoA long chain synthetase (*Acsl*) multigene family indicates that the two groups *Acsl3* and *Acsl4* (*Acsl3/4*) and *Acsl1 Acsl5* and *Acsl6* (*Acsl1/5/6*) arose from the duplication of two ancestral genes with the 2R WGD in the vertebrate ancestor (Chapter III.2 (Lopes-Marques *et al.*, 2013)). Similarly to *Acss,* differential paralogue retention uncovered an uncharacterized *Ascl* named *Acsl2* in the teleost lineage which was found to be paralogous to the *Acsl1/5/6* group. Additionally, the *Acsl* gene set in teleost was further expanded by the preservation 3R WGD duplicates.

Overall, a trend in the preservation of additional copies 2R or/and 3R WGD of *Acss* and *Acsl* genes in the teleost lineage is observed (Chapter III (Castro *et al.*, 2012b; Lopes-Marques *et al.*, 2013)). Generally, the preservation of duplicate redundant genes with overlapping gene expression which is the case for the *acss1a, b; acsl3a, b* and *acsl4a, b* is often observed when the corresponding transcript, in this case Acyl-CoA synthetases, is in high demand (Fig. 2A (Zhang, 2003)). Here, regarding the teleost lineage one can hypothesize that the preservation Acyl-CoA synthetases duplicates is a way to fulfill the high demand of FA activation for β-oxidation, given that FA oxidation is considered to be the main energy source in teleosts (Tocher, 2003). This observation may also extend to the perseveration of *acss1b* in migratory birds, given that an upregulation of FA transporters and carnitine palmitoyl transferase in flight muscles has previously been documented and linked to the use of FA as fuel in β-oxidation during migrations (Guglielmo *et al.*, 2002; McFarlan *et al.*, 2009; Guglielmo, 2010).

### VII.2 FATTY ACID BIOSYNTHESIS

The synthesis of LC-PUFAs from dietary EFAs requires the combined action of two distinct enzyme families ELOVL and FADS. The co-evolution of these two enzyme families in vertebrate species has been shaped by events of gene duplication, gene loss and diet, thus the ability to efficiently endogenously synthesize LC-PUFAS varies among vertebrate species (Castro *et al*., 2016). Ironically, although fish are the major source of LC-PUFAS in the human diet, many fish species are unable to efficiently complete the LC-PUFA biosynthesis endogenously, relying on dietary supplementation, due to possessing an incomplete enzyme set for LC-PUFA biosynthesis (Castro *et al*., 2012c; Castro *et al*., 2016). In aquaculture, this is overcome by the supplementation of the diets with fish oils and fish meal. Yet, these supplements imply serious limitations such as high cost of production to obtain relatively low values of DHA and EFA and environmental costs (Sidhu, 2003; Foran *et al*., 2005; Park *et al*., 2006; Tocher, 2010; Abedi *et al*., 2014). Thus, an enormous effort has been put into understanding the lipid metabolism in fish. Here research has focused on identification of the FA requirements and profile of several species, as well as functional characterization of the Elovl and Fads enzymes and the exploration for alternative sources of EFA (Tocher, 2003; Tocher, 2010; Tocher *et al*., 2015).

Previous studies have attempted to clarify the evolutionary history of the *Elovl* and *Fads* gene families and combine this information with enzyme functionality and diet. Yet, an exhaustive characterization of the genetic complement in species with key phylogenetic positioning remained incomplete. To amend this, the evolutionary history of the elongase enzymes and desaturase enzymes was investigated in several vertebrate lineages and functional characterization of several Elovl and Fads enzymes was performed (Chapter IV).

First, regarding the distribution of elongases namely *Elovl2* and *Elovl5*, phylogenetic and synteny analysis strongly suggests that they emerged from one gene in the vertebrate ancestor as a result of the 2R WGD (Fig. 2B, Chapter IV (Monroig *et al*., 2016)). Functional characterization of elongases isolated from key chordate species such as: *P. marinus*, *C. milii* and *B. lanceolatum,* suggests that functional diversification of the Elovl enzymes occurred after the 2R WGD (Chapter IV (Monroig *et al*., 2016)). An alternative scenario would involve the acquisition of a $C_{18}$ to $C_{22}$ elongation capacity in vertebrate ancestry, with the loss of $C_{22}$ elongation in the cyclostomes occurring after their divergence. Even so, the inability to elongate $C_{22}$ to $C_{24}$ may have less significant consequences in lampreys, possibly due to their parasitic diet as adults, which provides a direct access to LC-PUFAS. In fact, it has been shown that lampreys favored *Salvelinus namaycush* with higher fat content also known as siscowets instead of leans (lower fat content) (Goetz *et al.*, 2016). It has been postulated that this preference is possibly due to the siscowet high lipid content being more capable of energetically sustaining parasitism (Goetz *et al.*, 2016). Nonetheless, how larval lampreys access these LC-PUFAS remains unresolved. Furthermore, functional characterization of the *C. milii* elongases reveals that the efficient completion of endogenous LC-PUFA biosynthesis and the synthesis of DHA via the Sprecher pathway probably occurred for the first time in the basal gnathostomes (Sprecher, 2000; Monroig *et al.*, 2016).

Although the investigation of the evolutionary history of *Elovls* indicates that the functional diversification of these enzymes took place after the 2R WGD, to specify the time frame in which a full LC-PUFA biosynthesis pathway emerged one must also consider the evolutionary history of the *Fads* gene family. Previous research of the *Fads* gene family in vertebrates suggested that *Fads1* and *Fads2* emerged prior to gnathostome radiation, despite the fact that *Fads1* is absent in many actinopterygii species (Fig. 2B (Castro *et al.*, 2012c)). Here in Chapter IV.2 and IV.3, the analysis of the genetic repertoire of species placed in key phylogenetic positions supports these findings with the identification of *Fads1* and *Fads2* in the agnathan *L. japonicum*.

Furthermore, to clarify the timing of the *Fads1* loss, several actinopterygii lineages were investigated revealing the retention of *Fads1* in lineages that diverged prior to the 3R WGD (holostei, polypteriformes) and in one lineage (elopomorpha) that diverged after the 3R WGD (Chapter IV.3). This indicates that *Fads1* loss most probably only took place after the divergence of the elopomorpha lineage. Interestingly, two *Fads2* genes were found in several osteoglossomorpha species (with the exception of *A. gigas* for which no genome data is yet available), although phylogenetic positioning suggests that these genes correspond to 3R WGD retained duplicates, synteny analysis was unclear in supporting this hypothesis (Chapter IV.3). Functional analysis of *P. buchholzi Fads2* genes showed that *Fad2a* retained a phenotype to that observed in the human *Fads2* - Δ6/Δ8 activity, while the *Fads2b* displayed an alternative activity typically observed in *Fads1* - Δ5 activity (Chapter IV.3). In teleosts, the identification of Fads2 desaturases or multiple Fads2 desaturases, with alternative activities has mainly been

observed in marine herbivores and freshwater species (Li *et al*., 2010b; Tocher, 2010). The functional plasticity of *Fads2* observed in species with limited access to dietary LC-PUFAS has been suggested to be an evolutionary solution to overcome the bottleneck in LC-PUFA biosynthesis caused by the loss of *Fads1.* On the other hand, the loss of *Fads1* in marine species was suggested to imply less significant consequences in a DHA rich marine environment. Similarly to the gain of alternative desaturation activities by Fads2, it has been observed that in some teleost species the loss of *Elovl2* has been compensated by the gain of substrate preferences of Elovl4*.* For example, the Elovl4 in *Nibea mitsukurii* (Kabeya *et al*., 2015), *Siganus canaliculatus* (Monroig *et al*., 2012) and *Rachycentron canadum* (Monroig *et al*., 2011b) have been shown to be active towards $C_{20}$ and $C_{22}$ substrates.
Curiously, although the teleost lineage displays a rather variable *Elovl* and *Fads* gene repertoire, a common pattern of loss can be observed, with the preferential loss of *Elovl2* and *Fads1.* The favored retention of *Elovl5* and *Fad2* could partially be explained by their position in the LC-PUFA biosynthesis cascade. In effect, these two enzymes participate in the initial stages of the LC-PUFA pathway, processing dietary EFA. Therefore, the loss *Elovl5* and *Fad2* would imply the impairment of the whole LC-PUFA pathway, given that *Elovl2* and *Fads1* generally do not process EFA. Additionally, it has been shown that the functional plasticity observed in Fads desaturases is "*more prone*" to appear in Fads2 which requires a lower number of mutations to attain an alternative substrate specificity in comparison to Fads1 (Watanabe *et al*., 2016).

The LC-PUFA biosynthesis pathway is a result of a remarkable co-evolution of two distinct genes families *Elovl* and *Fads* with complementary roles. Intriguingly and although these gene families are unrelated, they both expanded at approximately the same time point in vertebrate evolution. While the expansion of *Elovl* gene family coincided with the 2R WGD, the expansion of the *Fads* gene family seems to be due to a tandem gene duplication that took place in the ancestral vertebrate. Expansion of both these gene families allowed for fine-tuning of enzymatic activities that eventually resulted in a complete LC-PUFA biosynthesis pathway. However, distinct genetic repertoires of *Elovl* and *Fads* are observed in vertebrates, with teleosts possibly being the most striking case with the observed loss of *Fads1* and *Elovl2* in numerous species.

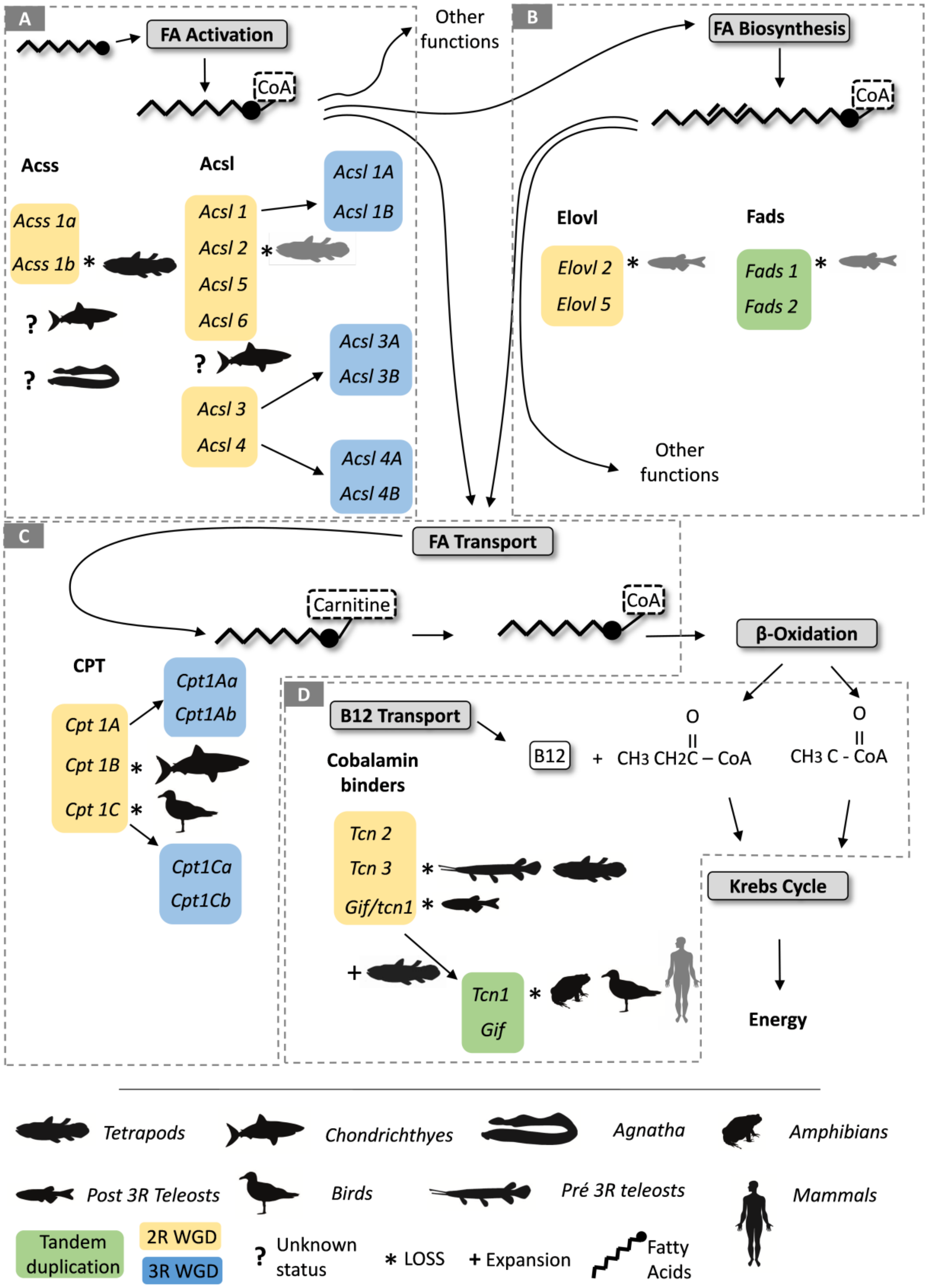


**Figure 2:** Schematic outline of the FA metabolic pathways covered in this thesis and the corresponding studied gene families. For each gene family, the general evolutionary history is depicted. Symbols and background colour indicate events or processes and animals silhouettes indicate the affected lineages. Grey silhouettes indicate that loss in the corresponding lineage does not affect all individuals.

## VII.3 β-OXIDATION

β-oxidation of long chain FA for energy production requires the transport of FA into the mitochondria. This transport is facilitated by the CPT enzymes (McGarry *et al.*, 1997). In mammals, three *CPT1* genes have been described: *CPT1A*, *CPT1B* and *CPT1C*. While orthologous of *CPT1A* and *CPT1B* have been documented in all major vertebrate lineages, *CPT1C* was considered to be mammalian specific (Wolfgang *et al.*, 2006; Boukouvala *et al.*, 2010; Lee *et al.*, 2012). The reexamination of the evolutionary history of this gene family revealed that although mammalian *CPT1C* presents a highly divergent sequence, its placing in the phylogenetic tree and genomic *locus* indicates that it is orthologous to teleost *cpt1a1* and not a mammalian novelty, thus being older than expected (Chapter V.1 (Lopes-Marques *et al.*, 2015)). Additionally, the expansion of the *CPT1* gene family was also found to be coincident with 2R WGD (Chapter V.1(Lopes-Marques *et al.*, 2015)). The analysis of evolutionary history uncovered the retention of distinct genetic sets of *Cpt1* genes in vertebrates (Fig. 2C), for example: *Cpt1c* seems to be absent from bird lineage, while *Cpt1b* is apparently lost in the chondrichthyes. Interestingly, the loss of the *Cpt1b,* the muscle isoform, in chondrichthyes correlates to the reported unusual energy metabolism observed in this lineage, who do not rely on β-oxidation but rather on ketones bodies as the main energy source (Speers-Roesch *et al.*, 2010). Whereas the consequences of the loss of *Cpt1c* are more challenging to infer, given that the known isoform in mammals is very divergent (Chapter V.1(Lopes-Marques *et al.*, 2015)). For example, in humans *Cpt1c* is brain-specific and localized to the endoplasmic reticulum, suggesting an alternative functional role, given that FA transport and β-oxidation do not fulfill the human brain energy requirements, which relies on glucose and ketone bodies as the main energy source (Lee *et al.*, 2012; Schönfeld *et al.*, 2013). Additionally, a larger set of *Cpt1* genes was identified in the teleost lineage due to the retention of  3R WGD duplicates (Fig. 2C, Chapter V.1 (Lopes-Marques *et al.*, 2015)). Curiously, this correlates with the retention of additional FA activation enzymes *Acss* and *Acsl* in the teleost lineage (Chapter III). Interestingly, the seemingly coordinated enrichment of the genetic repertoire in the pathways preceding β-oxidation (*Acss Acsl* and *Cpt1*) is in accordance with previous observations that indicate that β-oxidation of FA is the primary energy source in teleosts (Tocher, 2003).

Vitamin B12/cobalamin plays an critical role in the β-oxidation of unsaturated FAs and/or FAs presenting an odd number of carbons where it is essential in the conversion of propionyl-CoA to succinyl-CoA (Fig.2D (Lehninger *et al.*, 2008)). Previous research revealed contrasting gene repertoires in vertebrates, describing 3 cobalamin binders in human, 2 in mouse, birds and amphibians, and one binder in the teleost lineage, which presented intermediate characteristics to those observed in human cobalamin binders (Greibe *et al.*, 2012). These observations lead to the proposition that B12 binders diversified after the divergence of the teleost lineage (Greibe *et al.*, 2012). However, the re-examination of this gene family with the inclusion of *Tcn-like* sequences from chondrichthyes indicated otherwise (Chapter V.2 (Lopes-Marques *et al.*, 2015)). Still, the differential paralogue retention, gene loss and tandem gene duplication in tetrapods obscured the true evolutionary history. Database mining and phylogenetic analysis uncovered two uncharacterized binders (*Tcn3 and Tcn1/Gif*) in chondricthyes, which proved to be important for unraveling the evolutionary history of this gene family (Chapter V.2 (Lopes-Marques *et al.*, 2015)). When including the chondrichthyes binders phylogenetic and synteny analysis suggests that two independent events of expansion occurred in the cobalamin binder family: the first coincident with 2R WGD and the second in the tetrapod ancestor. The expansion of cobalamin binders in the tetrapod ancestor possibly allowed the sub-functionalization and fine tuning of these binders (*Tcn1, Tcn2* and *Gif*) to the distinct physiological compartments by adjusting specificity, affinity and resistance to digestion by digestive proteases. Curiously, the timing of cobalamin binder diversification in basal tetrapods concurs with the expansion of the pepsinogen gene family and the colonization of terrestrial habitats granting access to novel dietary sources.

### VII.4 Gastric proteases and protein digestion

Similarly to FA metabolism, protein metabolism has also been significantly modulated by gene/genome duplication, gene loss, diet and other environmental factors, resulting in distinct genetic repertoires and phenotypical outcomes (Ordoñez *et al.*, 2008; Castro *et al.*, 2014). Particularly, it has been shown that the genetic repertoire plays an important role in the elaboration of the digestive system. For example, the secondary loss of gastric glands has occurred several times in vertebrate evolution, and was found to be correlated with the presence or absence of a set of genes coding for the proton pump and pepsinogens (Ordoñez *et al.*, 2008; Castro *et al.*, 2014). Regarding the pepsinogens namely pepsinogen A, previous investigation revealed a variable genetic repertoire which was suggested to be a result of dietary adaptation, where higher levels of pepsinogens are generally found in animals with an herbivorous diet (Kageyama, 2002). Nevertheless, no attempt was made to elucidate the evolutionary history of the pepsinogen C (*PgC*) gene family. Initially, the *PgC* gene family was regarded as a single copy gene in vertebrate species (Kageyama, 2002). However, analysis of this gene family uncovered an unexpected diversity of *PgC* genes in several vertebrate lineages (Chapter VI.1 (Castro *et al.*, 2012a)). A highly variable assortment of gene sequences was identified in several vertebrates; ranging from species that present no *PgC* gene – e.g. the teleost *O. latipes*, to the marsupial *M. domestica* who presents 5 different *PgC* genes. Further phylogenetic and synteny analysis, suggested that the *PgC* diversification took place in basal tetrapod through tandem gene duplications (Chapter VI.1 (Castro *et al.*, 2012a)). Interestingly, similar to the observed in the cobalamin binders, the expansion of the *PgC* family coincides with the transition to terrestrial habitats and access to novel dietary sources. Thus, the expansion of the pepsinogens in the tetrapod ancestor was probably followed by the acquisition of alternative substrate preferences later observed in several studies (Kageyama, 2002; Narita *et al.*, 2002; Kageyama, 2006).

While in some lineages the acquisition of a larger set of pepsinogens posed most likely an advantage, the loss of pepsinogens, also very common, in vertebrate evolution signals events of dietary adaptation, which is apparently the case of Chymosin (*Cmy*). The investigation of the distribution of this gene family in mammals revealed an unforeseen independent number of gene loss events (Chapter VI.2). Previously, *Cmy* was documented as pseudogenized in human (Ord *et al.*, 1990). However, a closer investigation revealed novel cases of gene loss with at least 8 independent events of pseudogenization occurring in the 3 mammalian orders (cercopithecoidea, hystoricomoroha, perrissodactyla) and in several individual species (Chapter VI.2). Interestingly, Cmy, a neonatal protease, was previously correlated with immune transfer strategies and considered beneficial for passive immune transfer, due to its low proteolytic activity toward immunoglobulin-γ (IgG) (Foltmann, 1992; Kageyama, 2002; Baintner, 2007; Furukawa *et al.*, 2014). The combined analysis of the coding status of

*Cmy* with, immune transfer strategies in the various mammalian orders, reveals that *Cmy* loss seems to parallel the gain of maternal immune transfer (Chapter VI.2). Additionally, selection analysis shows that mammalian orders with a coding *Cmy* and presenting maternal transfer of IgG exhibited relaxed selection in *Cmy* gene. Here one can hypothesize that the most probable outcome for *Cmy* in these lineages is non-functionalization, or sub-functionalization as observed in the case of the owl monkey were *Cmy* is expressed in the adult stomach (Kageyama, 2000). The evolution of the *Cym* gene family represents an interesting case were the cross-talk between the immune system and digestive system seems to model the neonatal genetic repertoire of digestive enzymes.

# CHAPTER VIII

## FINAL REMARKS

# CHAPTER VIII – FINAL REMARKS

The analysis of the evolutionary history of several gene families intervening in FA metabolic pathways and protein digestion clearly shows that the resulting genetic repertoires observed in vertebrates is sculpted by events such as the 2R WGD, 3R WGD gene duplication, gene loss, mutation, life trajectory and diet among others.

The investigation of several gene families involved in FA metabolism in vertebrates visibly demonstrates the impact of the 2R WGD, with the expansion and diversification, of the *Acss1, Acsl, Elovl, Cpt1,* and cobalamin binder gene families. The teleost specific 3R WGD also played an important role in the expansion of the *Acsl* and *Cpt1* gene family. Conversely, when analyzing these gene families only in one case was the 4:1 ratio found (*Acsl1/2/5/6*), indicating that other events such as: gene loss also played a critical role in modeling the resulting genetic repertoires.

The colonization of terrestrial habitats by tetrapods constituted an important evolutionary event entailing many modifications, ranging from the development of limbs, to air breathing lungs and elaboration of the digestive system due to the access to novel dietary sources (Ashley-Ross *et al.*, 2013). In this work, it was possible to observe that the transition to terrestrial habitats and access to novel food sources can be correlated to the expansion of the cobalamin binder and pepsinogen C gene families. Additionally, diet and trophic level probably play an important role in modulating the genetic repertoire involved in the biosynthesis of LC-PUFAs, presenting cases of gene duplication/loss and functional plasticity observed in the teleost lineage. The observation of a larger genetic repertoire in the pathways preluding β-oxidation in teleost species was an interesting finding given that these species rely essentially in FA oxidation for energy provision. Finally, the investigation of the evolution of the *Cmy* gene family uncovered a new perspective where the cross-talk between digestive system and immune system apparently plays a decisive role regarding the Chymosin genetic repertoire in mammals.

# CHAPTER IX

## FUTURE DIRECTIONS

# CHAPTER IX – FUTURE DIRECTIONS

Taking into consideration the findings in the present work, an inevitable future direction should be to complete the characterization of the remaining families involved in FA metabolism in vertebrates. For example, the investigation of the evolutionary history of gene families that prelude β-oxidation would pose an interesting case which could further confirm the enrichment of this pathway in the teleost lineage and possibly uncover other species with a preferential use of FAs as an energy source. Additionally, it would also be interesting to verify if the transition to terrestrial habitats during vertebrate evolution also impacted the evolution of the digestive lipases, similarly to that observed in protein digestion. Furthermore, in the current "*Omics*" Era, a detailed characterization through transcriptomics of the genetic machinery involved in FA metabolism, in distinct tissues and/or developmental stages, or in processing distinct diets, would bring a more precise vision of these metabolic pathways.

Even so, many uncharted paths and pathways still remain to be explored.

# X

# REFERENCES

# X - REFERENCES

Abedi, E. and M. A. Sahari; 2014; "Long-chain polyunsaturated fatty acid sources and evaluation of their nutritional and functional properties." Food Science & Nutrition; 2;(5); 443-463.

Albalat, R. and C. Canestro; 2016; "Evolution by gene loss." Nat Rev Genet; 17;(7); 379-391.

Amores, A., J. Catchen, A. Ferrara, et al.; 2011; "Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication." Genetics; 188;(4); 799-808.

Amores, A., A. Force, Y.-L. Yan, et al.; 1998; "Zebrafish hox Clusters and Vertebrate Genome Evolution." Science; 282;(5394); 1711-1714.

Ashley-Ross, M. A., S. T. Hsieh, A. C. Gibb, et al.; 2013; "Vertebrate Land Invasions–Past, Present, and Future: An Introduction to the Symposium." Integrative and Comparative Biology; 53;(2); 192-196.

Avery, O. T., C. M. MacLeod and M. McCarty; 1944; "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types." The Journal of Experimental Medicine; 79;(2); 137-158.

Axelsson, E., A. Ratnakumar, M.-L. Arendt, et al.; 2013; "The genomic signature of dog domestication reveals adaptation to a starch-rich diet." Nature; 495;(7441); 360-364.

Baintner, K.; 2007; "Transmission of antibodies from mother to young: Evolutionary strategies in a proteolytic environment." Veterinary Immunology and Immunopathology; 117;(3–4); 153-161.

Blanchard, H., P. Legrand and F. Pédrono; 2011; "Fatty Acid Desaturase 3 (*Fads3*) is a singular member of the Fads cluster." Biochimie; 93;(1); 87-90.

Bonnefont, J.-P., F. Djouadi, C. Prip-Buus, et al.; 2004; "Carnitine palmitoyltransferases 1 and 2: biochemical, molecular and medical aspects." Molecular Aspects of Medicine; 25;(5–6); 495-520.

Boukouvala, E., M. J. Leaver, L. Favre-Krey, et al.; 2010; "Molecular characterization of a gilthead sea bream (Sparus aurata) muscle tissue cDNA for carnitine palmitoyltransferase 1B (CPT1B)." Comparative Biochemistry and Physiology - B Biochemistry and Molecular Biology; 157;(2); 189-197.

Braasch, I., A. R. Gehrke, J. J. Smith, et al.; 2016; "The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons." Nat Genet; 48;(4); 427-437.

Briani, C., C. Dalla Torre, V. Citton, et al.; 2013; "Cobalamin Deficiency: Clinical Picture and Radiological Findings." Nutrients; 5;(11); 4521-4539.

Bridges, C. B.; 1936; "The bar "gene" a duplication." Science; 83;(2148); 210-211.

Byrne, C. S., E. S. Chambers, D. J. Morrison, et al.; 2015; "The role of short chain fatty acids in appetite regulation and energy homeostasis." International Journal of Obesity (2005); 39;(9); 1331-1338.

Cañestro, C. (2012). Two Rounds of Whole-Genome Duplication: Evidence and Impact on the Evolution of Vertebrate Innovations. Polyploidy and Genome Evolution. P. S. Soltis and D. E. Soltis. Berlin, Heidelberg, Springer Berlin Heidelberg: 309-339.

Canfora, E. E., J. W. Jocken and E. E. Blaak; 2015; "Short-chain fatty acids in control of body weight and insulin sensitivity." Nat Rev Endocrinol; 11;(10); 577-591.

Carginale, V., F. Trinchella, C. Capasso, et al.; 2004; "Adaptive evolution and functional divergence of pepsin gene family." Gene; 333;(81-90.

Castro, L. F. C., O. Gonçalves, S. Mazan, et al.; 2014; "Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history." Proceedings of the Royal Society B: Biological Sciences; 281;(1775);

Castro, L. F. C., M. Lopes-Marques, O. Gonçalves, et al.; 2012a; "The Evolution of Pepsinogen C Genes in Vertebrates: Duplication, Loss and Functional Diversification." PLoS One; 7;(3); e32852.

Castro, L. F. C., M. Lopes-Marques, J. M. Wilson, et al.; 2012b; "A novel Acetyl-CoA synthetase short-chain subfamily member 1 (*Acss1*) gene indicates a dynamic history of paralogue retention and loss in vertebrates." Gene; 497;(2); 249-255.

Castro, L. F. C., Ó. Monroig, M. J. Leaver, et al.; 2012c; "Functional Desaturase Fads1 (Δ5) and Fads2 (Δ6) Orthologues Evolved before the Origin of Jawed Vertebrates." PLoS ONE; 7;(2); e31950.

Castro, L. F. C., D. R. Tocher and O. Monroig; 2016; "Long-chain polyunsaturated fatty acid biosynthesis in chordates: Insights into the evolution of *Fads* and *Elovl* gene repertoire." Progress in Lipid Research; 62;(25-40.

Chan, A. S., M. H. Horn, K. A. Dickson, et al.; 2004; "Digestive enzyme activities in carnivores and herbivores: comparisons among four closely related prickleback fishes (Teleostei: Stichaeidae) from a California rocky intertidal habitat." Journal of Fish Biology; 65;(3); 848-858.

Chen, J.-N., S. Samadi and W.-J. Chen (2015). Elopomorpha (Teleostei) as a New Model Fish Group for Evolutionary Biology and Comparative Genomics. Evolutionary Biology: Biodiversification from Genotype to Phenotype. P. Pontarotti. Cham, Springer International Publishing: 329-344.

Chen, S., B. H. Krinsky and M. Long; 2013; "New genes as drivers of phenotypic evolution." Nat Rev Genet; 14;(9); 645-660.

Cliften, P. F., R. S. Fulton, R. K. Wilson, et al.; 2006; "After the Duplication: Gene Loss and Adaptation in Saccharomyces Genomes." Genetics; 172;(2); 863-872.

Cole, N. J. and P. D. Currie; 2007; "Insights from sharks: Evolutionary and developmental models of fin development." Developmental Dynamics; 236;(9); 2421-2431.

Collins, S. A., G. Sinclair, S. McIntosh, et al.; 2010; "Carnitine palmitoyltransferase 1A (CPT1A) P479L prevalence in live newborns in Yukon, Northwest Territories, and Nunavut." Molecular Genetics and Metabolism; 101;(2); 200-204.

Cox, T. M.; 1999; "Mendel and his legacy." QJM: An International Journal of Medicine; 92;(4); 183-186.

Crick, F.; 1970; "Central Dogma of Molecular Biology." Nature; 227;(5258); 561-563.

Crow, K. D., C. D. Smith, J.-F. Cheng, et al.; 2012; "An Independent Genome Duplication Inferred from Hox Paralogs in the American Paddlefish—A Representative Basal Ray-Finned Fish and Important Comparative Reference." Genome Biology and Evolution; 4;(9); 937-953.

Darwin, C.; 1859; The Origen of Species by the means of Natural Selection or the Preservation of the Favored Races in the Struggle for Life. London, John Murray, Albemarle street.

Darwin, C.; 1887; Charles Darwin´s Autobiography London, Bibliolis Books.

Dean, M., M. Carrington, C. Winkler, et al.; 1996; "Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the *CKR5* Structural Gene." Science; 273;(5283); 1856-1862.

Dehal, P. and J. L. Boore; 2005; "Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate." PLOS Biology; 3;(10); e314.

Docker., M. F., J. B. Hume and B. J. Clemens (2015). Introduction: A Surfeit of Lampreys. Lampreys: Biology, Conservation and Control. Margaret F. Docker, Springer **37:** 438.

Dunning, K. R., M. R. Anastasi, V. J. Zhang, et al.; 2014; "Regulation of Fatty Acid Oxidation in Mouse Cumulus-Oocyte Complexes during Maturation and Modulation by PPAR Agonists." PLoS One; 9;(2); e87327.

Emerling, C. A. and M. S. Springer; 2014; "Eyes underground: Regression of visual protein networks in subterranean mammals." Molecular Phylogenetics and Evolution; 78;(260-270.

Esser, V., N. F. Brown, A. T. Cowan, et al.; 1996; "Expression of a cDNA isolated from rat brown adipose tissue and heart identifies the product as the muscle isoform of carnitine palmitoyltransferase I (M-CPT I): M-CPT I is the predominant CPT I isoform expressed in both white (epididymal) and brown adipocytes." Journal of Biological Chemistry; 271;(12); 6972-6977.

Fedosov, S. N., N. U. Fedosova, B. Kräutler, et al.; 2007; "Mechanisms of Discrimination between Cobalamins and Their Natural Analogues during Their Binding to the Specific B12-Transporting Proteins." Biochemistry; 46;(21); 6446-6458.

Finegold, L.; 1986; "Molecular aspects of adaptation to extreme cold environments." Advances in Space Research; 6;(12); 257-264.

Foltmann, B.; 1992; "Chymosin: a short review on foetal and neonatal gastric proteases." Scand J Clin Lab Invest Suppl; 210;(65-79.

Foran, J. A., D. H. Good, D. O. Carpenter, et al.; 2005; "Quantitative Analysis of the Benefits and Risks of Consuming Farmed and Wild Salmon." The Journal of Nutrition; 135;(11); 2639-2643.

Force, A., M. Lynch, F. B. Pickett, et al.; 1999; "Preservation of duplicate genes by complementary, degenerative mutations." Genetics; 151;(

Freitas, R., G. Zhang and M. J. Cohn; 2006; "Evidence that mechanisms of fin development evolved in the midline of early vertebrates." Nature; 442;(7106); 1033-1037.

Fucho, R., N. Casals, D. Serra, et al.; 2016; "Ceramides and mitochondrial fatty acid oxidation in obesity." The FASEB Journal; fj. 201601156R.

Fumagalli, M., I. Moltke, N. Grarup, et al.; 2015; "Greenlandic Inuit show genetic signatures of diet and climate adaptation." Science; 349;(6254); 1343-1347.

Furukawa, S., Y. Kuroda and A. Sugiyama; 2014; "A Comparison of the Histological Structure of the Placenta in Experimental Animals." Journal of Toxicologic Pathology; 27;(1); 11-18.

Gallardo, M. H., J. W. Bickham, R. L. Honeycutt, et al.; 1999; "Discovery of tetraploidy in a mammal." Nature; 401;(6751); 341-341.

Gess, R. W., M. I. Coates and B. S. Rubidge; 2006; "A lamprey from the Devonian period of South Africa." Nature; 443;(7114); 981-984.

Glasauer, S. M. K. and S. C. F. Neuhauss; 2014; "Whole-genome duplication in teleost fishes and its evolutionary consequences." Molecular Genetics and Genomics; 289;(6); 1045-1060.

Goetz, F., S. E. Smith, G. Goetz, et al.; 2016; "Sea lampreys elicit strong transcriptomic responses in the lake trout liver during parasitism." BMC Genomics; 17;(1); 675.

Grafflin, A. L. and D. Green; 1948; "Studies on the cyclophorase system II. The complete oxidation of fatty acids." Journal of Biological Chemistry; 176;(1); 95-115.

Greibe, E., S. Fedosov and E. Nexo; 2012; "The Cobalamin-Binding Protein in Zebrafish Is an Intermediate between the Three Cobalamin-Binding Proteins in Human." PLoS One; 7;(4); e35660.

Griffith, F.; 1928; "The Significance of Pneumococcal Types." The Journal of Hygiene; 27;(2); 113-159.

Grygiel-Górniak, B.; 2014; "Peroxisome proliferator-activated receptors and their ligands: nutritional and clinical implications - a review." Nutrition Journal; 13;(1); 17.

Guglielmo, C. G.; 2010; "Move That Fatty Acid: Fuel Selection and Transport in Migratory Birds and Bats." Integrative and Comparative Biology; 50;(3); 336-345.

Guglielmo, C. G., N. H. Haunerland, P. W. Hochachka, et al.; 2002; "Seasonal dynamics of flight muscle fatty acid binding protein and catabolic enzymes in a migratory shorebird." American Journal of Physiology - Regulatory, Integrative and Comparative Physiology; 282;(5); R1405-R1413.

Guillou, H., D. Zadravec, P. G. P. Martin, et al.; 2010; "The key roles of elongases and desaturases in mammalian fatty acid metabolism: Insights from transgenic mice." Progress in Lipid Research; 49;(2); 186-199.

Hassam, A. G., J. P. W. Rivers and M. A. Crawford; 1977; "The Failure of the Cat to Desaturate Linoleic Acid; Its Nutritional Implications." Annals of Nutrition and Metabolism; 21;(5); 321-328.

Hastings, N., M. Agaba, D. R. Tocher, et al.; 2001; "A vertebrate fatty acid desaturase with Δ5 and Δ6 activities." Proceedings of the National Academy of Sciences; 98;(25); 14304-14309.

Hedges, S. B., J. Dudley and S. Kumar; 2006; "TimeTree: a public knowledge-base of divergence times among organisms." Bioinformatics; 22;(23); 2971-2972.

Hedges, S. B., J. Marin, M. Suleski, et al.; 2015; "Tree of Life Reveals Clock-Like Speciation and Diversification." Molecular Biology and Evolution; 32;(4); 835-845.

Henkel, C. V., E. Burgerhout, D. L. de Wijze, et al.; 2012a; "Primitive Duplicate Hox Clusters in the European Eel's Genome." PLoS One; 7;(2); e32231.

Henkel, C. V., R. P. Dirks, D. L. de Wijze, et al.; 2012b; "First draft genome sequence of the Japanese eel, Anguilla japonica." Gene; 511;(2); 195-201.

Higgs, P. G. and T. K. Attwood (2004). Phylogenetic Methods. Bioinformatics and Molecular Evolution, Blackwell Publishing Ltd.: 158-194.

Hoegg, S. and A. Meyer; 2005; "Hox clusters as models for vertebrate genome evolution." Trends in Genetics; 21;(8); 421-424.

Hokamp, K., A. McLysaght and K. H. Wolfe; 2003; "The 2R hypothesis and the human genome sequence." Journal of Structural and Functional Genomics; 3;(1); 95-110.

Holland, L. Z., R. Albalat, K. Azumi, et al.; 2008; "The amphioxus genome illuminates vertebrate origins and cephalochordate biology." Genome Research; 18;(7); 1100-1111.

Holland, L. Z., V. Laudet and M. Schubert; 2004; "The chordate amphioxus: an emerging model organism for developmental biology." Cellular and Molecular Life Sciences CMLS; 61;(18); 2290-2308.

Holland, P.; 1992; "Problems and paradigms: Hoemeobox genes in vertebrate evolution." BioEssays; 14;(4); 267-273.

Holland, P. W. H.; 2013; "Evolution of homeobox genes." Wiley Interdisciplinary Reviews: Developmental Biology; 2;(1); 31-45.

Holland, P. W. H., J. Garcia-Fernàndez, N. A. Williams, et al.; 1994; "Gene duplications and the origins of vertebrate development." Development; 1994;(Supplement); 125-133.

Huang, S., H. Tian, Z. Chen, et al.; 2010; "The evolution of vertebrate tetraspanins: gene loss, retention, and massive positive selection after whole genome duplications." BMC Evolutionary Biology; 10;(1); 306.

Hughes, A. L.; 1999; "Phylogenies of Developmentally Important Proteins Do Not Support the Hypothesis of Two Rounds of Genome Duplication Early in Vertebrate History." Journal of Molecular Evolution; 48;(5); 565-576.

Hughes, A. L. and R. Friedman; 2003; "2R or not 2R: Testing hypotheses of genome duplication in early vertebrates." Journal of Structural and Functional Genomics; 3;(1); 85-93.

Hughes, M. K. and A. L. Hughes; 1993; "Evolution of duplicate genes in a tetraploid animal, Xenopus laevis." Molecular Biology and Evolution; 10;(6); 1360-1369.

Innan, H. and F. Kondrashov; 2010; "The evolution of gene duplications: classifying and distinguishing between models." Nat Rev Genet; 11;(2); 97-108.

Inoue, J., Y. Sato, R. Sinclair, et al.; 2015; "Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling." Proceedings of the National Academy of Sciences; 112;(48); 14918-14923.

Jaillon, O., J.-M. Aury, F. Brunet, et al.; 2004; "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype." Nature; 431;(7011); 946-957.

Jeffery, W. R.; 2009; "Regressive Evolution in Astyanax Cavefish." Annual Review of Genetics; 43;(1); 25-47.

Jiang, Y., M. Xie, W. Chen, et al.; 2014; "The sheep genome illuminates biology of the rumen and lipid metabolism." Science; 344;(6188); 1168-1173.

Kabeya, N., Y. Yamamoto, S. F. Cummins, et al.; 2015; "Polyunsaturated fatty acid metabolism in a marine teleost, Nibe croaker Nibea mitsukurii: Functional characterization of Fads2 desaturase and Elovl5 and Elovl4 elongases." Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology; 188;(37-45.

Kageyama, T.; 2000; "New world monkey pepsinogens A and C, and prochymosins. Purification, characterization of enzymatic properties, cDNA cloning, and molecular evolution." J Biochem; 127;(5); 761-770.

Kageyama, T.; 2002; "Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development." Cellular and Molecular Life Sciences CMLS; 59;(2); 288-306.

Kageyama, T.; 2006; "Roles of Tyr13 and Phe219 in the unique substrate specificity of pepsin B." Biochemistry; 45;(48); 14415-14426.

Kao, Y.-H., J. H. Youson, J. A. Holmes, et al.; 1997a; "Changes in lipolysis and lipogenesis in selected tissues of the landlocked lamprey, Petromyzon marinus, during metamorphosis." Journal of Experimental Zoology; 277;(4); 301-312.

Kao, Y.-h., J. H. Youson and M. A. Sheridan; 1997b; "Differences in the total lipid and lipid class composition of larvae and metamorphosing sea lampreys, Petromyzon marinus." Fish Physiology and Biochemistry; 16;(4); 281-290.

Kim, E. B., X. Fang, A. A. Fushan, et al.; 2011; "Genome sequencing reveals insights into physiology and longevity of the naked mole rat." Nature; 479;(7372); 223-227.

Knox, W. E., B. Noyce and V. Auerbach; 1948; "Studies on the cyclophorase system; obligatory sparking of fatty acid oxidation." Journal of Biological Chemistry; 176;(1); 117-122.

Kocandrle, R. and K. Kleisner; 2013; "Evolution Born of Moisture: Analogies and Parallels Between Anaximander’s Ideas on Origin of Life and Man and Later Pre-Darwinian and Darwinian Evolutionary Concepts." Journal of the History of Biology; 46;(1); 103-124.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, et al.; 2002; "Selection in the evolution of gene duplications." Genome Biology; 3;(2); research0008.0001-research0008.0009.

Kothapalli, K. S. D., Ye, Kaixiong, M. S. Gadgil, et al.; 2016; "Positive Selection on a Regulatory Insertion–Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis of Arachidonic Acid." Molecular Biology and Evolution; 33;(7); 1726-1739.

Kuraku, S.; 2013; "Impact of asymmetric gene repertoire between cyclostomes and gnathostomes." Seminars in Cell & Developmental Biology; 24;(2); 119-127.

Kuraku, S., A. Meyer and S. Kuratani; 2009; "Timing of Genome Duplications Relative to the Origin of the Vertebrates: Did Cyclostomes Diverge before or after?"; Molecular Biology and Evolution; 26;(1); 47-59.

Kurokawa, T., S. Uji and T. Suzuki; 2005; "Identification of pepsinogen gene in the genome of stomachless fish, Takifugu rubripes." Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology; 140;(1); 133-140.

Lee, J. and M. Wolfgang; 2012; "Metabolomic profiling reveals a role for CPT1c in neuronal oxidative metabolism." BMC Biochemistry; 13;(1); 23.

Lee, P., S. Smith, J. Linderman, et al.; 2014; "Temperature-Acclimated Brown Adipose Tissue Modulates Insulin Sensitivity in Humans." Diabetes; 63;(11); 3686-3698.

Lehninger, A., D. Nelson and M. Cox; 2008; Lehninger Principles of Biochemistry, W. H. Freeman.

Lemey, P.; 2009; The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing, Cambridge University Press.

Leonard, A. E., B. Kelder, E. G. Bobik, et al.; 2002; "Identification and expression of mammalian long-chain PUFA elongation enzymes." Lipids; 37;(8); 733-740.

Leonard, A. E., S. L. Pereira, H. Sprecher, et al.; 2004; "Elongation of long-chain fatty acids." Progress in Lipid Research; 43;(1); 36-54.

Li, L. O., E. L. Klett and R. A. Coleman; 2010a; "Acyl-CoA synthesis, lipid metabolism and lipotoxicity." Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids; 1801;(3); 246-251.

Li, Y., O. Monroig, L. Zhang, et al.; 2010b; "Vertebrate fatty acyl desaturase with Δ4 activity." Proceedings of the National Academy of Sciences; 107;(39); 16840-16845.

Linnaeus, C.; 1735; Systema naturae, sive regna tria naturae systematice proposita per classes, ordines, genera, & species. Lugduni Batavorum, Theodorum Haak.

Lopes-Marques, M., I. Cunha, M. A. Reis-Henriques, et al.; 2013; "Diversity and history of the long-chain acyl-CoA synthetase (Acsl) gene family in vertebrates." BMC Evolutionary Biology; 13;(1); 271.

Lopes-Marques, M., I. L. S. Delgado, R. Ruivo, et al.; 2015; "The Origin and Diversity of Cpt1 Genes in Vertebrate Species." PLoS One; 10;(9); e0138447.

Los, D. A. and N. Murata; 1998; "Structure and expression of fatty acid desaturases." Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism; 1394;(1); 3-15.

Louis, E. J.; 2007; "Evolutionary genetics: Making the most of redundancy." Nature; 449;(7163); 673-674.

Lowe, D. R., F. W. H. Beamish and I. C. Potter; 1973; "Changes in the proximate body composition of the landlocked sea lamprey Petromyzon marinus (L.) during larval life and metamorphosis." Journal of Fish Biology; 5;(6); 673-682.

Lundin, L.-G., D. Larhammar and F. Hallböök; 2003; "Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates." Journal of Structural and Functional Genomics; 3;(1); 53-63.

Lynch, M. and J. S. Conery; 2000; "The Evolutionary Fate and Consequences of Duplicate Genes." Science; 290;(5494); 1151-1155.

MacCarthy, T. and A. Bergman; 2007; "The limits of subfunctionalization." BMC Evolutionary Biology; 7;(1); 213.

Macqueen, D. J. and I. A. Johnston; 2014; "A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification." Proceedings of the Royal Society B: Biological Sciences; 281;(1778);

Marquardt, A., H. Stöhr, K. White, et al.; 2000; "cDNA Cloning, Genomic Structure, and Chromosomal Localization of Three Members of the Human Fatty Acid Desaturase Family." Genomics; 66;(2); 175-183.

Martin, A.; 2001; "Is Tetralogy True? Lack of Support for the “One-to-Four Rule”." Molecular Biology and Evolution; 18;(1); 89-93.

Martin, K. J. and P. W. H. Holland; 2014; "Enigmatic Orthology Relationships between Hox Clusters of the African Butterfly Fish and Other Teleosts Following Ancient Whole-Genome Duplication." Molecular Biology and Evolution; 31;(10); 2592-2611.

Mathias, R. A., W. Fu, J. M. Akey, et al.; 2012; "Adaptive Evolution of the FADS Gene Cluster within Africa." PLoS One; 7;(9); e44926.

McCauley, D. W., M. F. Docker, S. Whyard, et al.; 2015; "Lampreys as Diverse Model Organisms in the Genomics Era." Bioscience; 65;(11); 1046-1056.

McFarlan, J. T., A. Bonen and C. G. Guglielmo; 2009; "Seasonal upregulation of fatty acid transporters in flight muscles of migratory white-throated sparrows (<em>Zonotrichia albicollis</em>)." Journal of Experimental Biology; 212;(18); 2934-2940.

McGarry, J. D. and N. F. Brown; 1997; "The Mitochondrial Carnitine Palmitoyltransferase System — From Concept to Molecular Analysis." European Journal of Biochemistry; 244;(1); 1-14.

Mehta, T. K., V. Ravi, S. Yamasaki, et al.; 2013; "Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum)." Proceedings of the National Academy of Sciences; 110;(40); 16044-16049.

Meyer, A. and M. Schartl; 1999; "Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions." Current Opinion in Cell Biology; 11;(6); 699-704.

Moghadam, H. K., M. M. Ferguson and R. G. Danzmann; 2011; "Whole genome duplication: challenges and considerations associated with sequence orthology assignment in Salmoninae." Journal of Fish Biology; 79;(3); 561-574.

Monroig, O., Y. Li and D. R. Tocher; 2011a; "Delta-8 desaturation activity varies among fatty acyl desaturases of teleost fish: high activity in delta-6 desaturases of marine species." Comp Biochem Physiol B Biochem Mol Biol; 159;(4); 206-213.

Monroig, Ó., M. Lopes-Marques, J. C. Navarro, et al.; 2016; "Evolutionary functional elaboration of the Elovl2/5 gene family in chordates." Scientific Reports; 6;(20510.

Monroig, Ó., S. Wang, L. Zhang, et al.; 2012; "Elongation of long-chain fatty acids in rabbitfish *Siganus canaliculatus*: Cloning, functional characterisation and tissue distribution of Elovl5- and Elovl4-like elongases." Aquaculture; 350–353;(63-70.

Monroig, Ó., K. Webb, L. Ibarra-Castro, et al.; 2011b; "Biosynthesis of long-chain polyunsaturated fatty acids in marine fish: Characterization of an Elovl4-like elongase from cobia Rachycentron canadum and activation of the pathway during early life stages." Aquaculture; 312;(1–4); 145-153.

Morais, S., O. Monroig, X. Zheng, et al.; 2009; "Highly Unsaturated Fatty Acid Synthesis in Atlantic Salmon: Characterization of ELOVL5- and ELOVL2-like Elongases." Marine Biotechnology; 11;(5); 627-639.

Moran, D., R. Softley and E. J. Warrant; 2015; "The energetic cost of vision and the evolution of eyeless Mexican cavefish." Science Advances; 1;(8);

Morino, K., K. F. Petersen and G. I. Shulman (2006). Molecular mechanisms of insulin resistance in humans and their potential links with mitochondrial dysfunction, Am Diabetes Assoc.

Morrison, D. J. and T. Preston; 2016; "Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism." Gut Microbes; 7;(3); 189-200.

Nakatani, Y., H. Takeda, Y. Kohara, et al.; 2007; "Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates." Genome Research; 17;(9); 1254-1265.

Narita, Y., S.-i. Oda, O. Takenaka, et al.; 2010; "Lineage-Specific Duplication and Loss of Pepsinogen Genes in Hominoid Evolution." Journal of Molecular Evolution; 70;(4); 313-324.

Narita, Y., S. Oda, A. Moriyama, et al.; 2002; "Primary structure, unique enzymatic properties, and molecular evolution of pepsinogen B and pepsin B." Arch Biochem Biophys; 404;(2); 177-185.

Nery, M. F., J. I. Arroyo and J. C. Opazo; 2014; "Increased rate of hair keratin gene loss in the cetacean lineage." BMC Genomics; 15;(1); 869.

Nirenberg, M. W. and J. H. Matthaei; 1961; "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides."; Proceedings of the National Academy of Sciences of the United States of America; 47;(10); 1588-1602.

Oh, J. W., O. Chung, Y. S. Cho, et al.; 2015; "Gene loss in keratinization programs accompanies adaptation of Cetacean skin to aquatic lifestyle." Experimental dermatology; 24;(8); 572-573.

Ohno, S.; 1970; Evolution by duplicaton New York, Springer Science + Business Media.

Ohno, S.; 2013; Evolution by gene duplication, Springer Science & Business Media.

Ohno, S., U. Wolf and N. B. Atkin; 1968; "Evolution from fish to mammals by gene duplication." Hereditas; 59;(1); 169-187.

Olson, M. V.; 1999; "When Less Is More: Gene Loss as an Engine of Evolutionary Change." The American Journal of Human Genetics; 64;(1); 18-23.

Ord, T., M. Kolmer, R. Villems, et al.; 1990; "Structure of the human genomic region homologous to the bovine prochymosin-encoding gene." Gene; 91;(2); 241-246.

Ordoñez, G. R., L. W. Hillier, W. C. Warren, et al.; 2008; "Loss of genes implicated in gastric function during platypus evolution." Genome Biology; 9;(5); R81-R81.

Panopoulou, G., S. Hennig, D. Groth, et al.; 2003; "New Evidence for Genome-Wide Duplications at the Origin of Vertebrates Using an Amphioxus Gene Set and Completed Animal Genomes." Genome Research; 13;(6a); 1056-1066.

Park, S. and M. A. Johnson; 2006; "Awareness of Fish Advisories and Mercury Exposure in Women of Childbearing Age." Nutrition Reviews; 64;(5); 250-256.

Perry, G. H., N. J. Dominy, K. G. Claw, et al.; 2007; "Diet and the evolution of human amylase gene copy number variation." Nature genetics; 39;(10); 1256-1260.

Pethybridge, H. R., C. C. Parrish, B. D. Bruce, et al.; 2014; "Lipid, Fatty Acid and Energy Density Profiles of White Sharks: Insights into the Feeding Ecology and Ecophysiology of a Complex Top Predator." PLoS One; 9;(5); e97877.

Putnam, N. H., T. Butts, D. E. K. Ferrier, et al.; 2008; "The amphioxus genome and the evolution of the chordate karyotype." Nature; 453;(7198); 1064-1071.

Qian, W., B.-Y. Liao, A. Y. F. Chang, et al.; 2010; "Maintenance of duplicate genes and their functional redundancy by reduced expression." Trends in genetics : TIG; 26;(10); 425-430.

Quadros, E. V.; 2010; "Advances in the Understanding of Cobalamin Assimilation and Metabolism." British journal of haematology; 148;(2); 195-204.

Read, T. D., R. A. Petit III, S. J. Joseph, et al. (2015). Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: Rhincodon typus Smith 1828, PeerJ PrePrints.

Richards, R. J.; 2009; Darwin's Place in the History of Thought: A Reevaluation. Washington DC, National Academies Press.

Richter, C., T. Tanaka and R. Y. Yada; 1998; "Mechanism of activation of the gastric aspartic proteinases: pepsinogen, progastricsin and prochymosin." Biochemical Journal; 335;(Pt 3); 481-490.

Rivers, J. P. W., A. J. Sinclair and M. A. Crawford; 1975; "Inability of the cat to desaturate essential fatty acids." Nature; 258;(5531); 171-173.

Roberts, R. B.; 1962; "Further implications of the doublet code."; Proceedings of the National Academy of Sciences of the United States of America; 48;(7); 1245-1250.

Robinson, L. E. and V. C. Mazurak; 2013; "N-3 Polyunsaturated Fatty Acids: Relationship to Inflammation in Healthy Adults and Adults Exhibiting Features of Metabolic Syndrome." Lipids; 48;(4); 319-332.

Röhrig, F. and A. Schulze; 2016; "The multifaceted roles of fatty acid synthesis in cancer." Nature Reviews Cancer;

Ruse, M.; 2009; The Darwinian Revolution:Rethinking Its Meaning and Significance. Washington DC, The National Academies press.

Schmitz, G. and J. Ecker; 2008; "The opposing effects of n−3 and n−6 fatty acids." Progress in Lipid Research; 47;(2); 147-155.

Schönfeld, P. and G. Reiser; 2013; "Why does brain metabolism not favor burning of fatty acids to provide energy? - Reflections on disadvantages of the use of free fatty acids as fuel for brain." Journal of Cerebral Blood Flow & Metabolism; 33;(10); 1493-1499.

Scienski, K., J. C. Fay and G. C. Conant; 2015; "Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex." Genome Biology and Evolution; 7;(12); 3249-3258.

Session, A. M., Y. Uno, T. Kwon, et al.; 2016; "Genome evolution in the allotetraploid frog Xenopus laevis." Nature; 538;(7625); 336-343.

Shimeld, S. M. and P. W. H. Holland; 2000; "Vertebrate innovations." Proceedings of the National Academy of Sciences; 97;(9); 4449-4452.

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.

Sidow, A.; 1996; "Gen(om)e duplications in the evolution of early vertebrates." Current Opinion in Genetics & Development; 6;(6); 715-722.

Skrabanek, L. and K. H. Wolfe; 1998; "Eukaryote genome duplication - where's the evidence?"; Current Opinion in Genetics & Development; 8;(6); 694-700.

Smith, D. M., B. T. Golding and L. Radom; 1999; "Understanding the Mechanism of B12-Dependent Methylmalonyl-CoA Mutase: Partial Proton Transfer in Action." Journal of the American Chemical Society; 121;(40); 9388-9399.

Smith, J. J. and M. C. Keinath; 2015; "The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications." Genome Research; 25;(8); 1081-1090.

Smith, J. J., S. Kuraku, C. Holt, et al.; 2013; "Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution." Nat Genet; 45;(4); 415-421.

Soupene, E. and F. A. Kuypers; 2008; "Mammalian Long-Chain Acyl-CoA Synthetases." Experimental Biology and Medicine; 233;(5); 507-521.

Speers-Roesch, B. and J. R. Treberg; 2010; "The unusual energy metabolism of elasmobranch fishes." Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology; 155;(4); 417-434.

Sprecher, H.; 2000; "Metabolism of highly unsaturated n-3 and n-6 fatty acids." Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids; 1486;(2–3); 219-231.

Stedman, H. H., B. W. Kozyak, A. Nelson, et al.; 2004; "Myosin gene mutation correlates with anatomical changes in the human lineage." Nature; 428;(6981); 415-418.

Sutton, W. S.; 1902; "On the Morphology of the Chromosome Group in Brachystola Magna." The Biological Bulletin; 4;(1); 24-39.

Sutton, W. S.; 1903; "The chromosomes in heredity." The Biological Bulletin; 4;(5); 231-250.

Tocher, D. R.; 2003; "Metabolism and Functions of Lipids and Fatty Acids in Teleost Fish." Reviews in Fisheries Science; 11;(2); 107-184.

Tocher, D. R.; 2010; "Fatty acid requirements in ontogeny of marine and freshwater fish." Aquaculture Research; 41;(5); 717-732.

Tocher, D. R. and B. D. Glencross (2015). Lipids and Fatty Acids. Dietary Nutrients, Additives, and Fish Health, John Wiley & Sons, Inc: 47-94.

Trevisanato, S.; 2016; "Reconstructing Anaximander´s Biological model unveils a theory of evolution akin to Darwins, though centuries before the birth of science."; Acta medico-historica Adriatica; 14;(1); 63-72.

Trevizan, L., A. de Mello Kessler, J. T. Brenna, et al.; 2012; "Maintenance of Arachidonic Acid and Evidence of Δ5 Desaturation in Cats Fed γ-Linolenic and Linoleic Acid Enriched Diets." Lipids; 47;(4); 413-423.

Van der Leij, F. R., N. C. A. Huijkman, C. Boomsma, et al.; 2000; "Genomics of the human carnitine acyltransferase genes." Molecular Genetics and Metabolism; 71;(1-2); 139-153.

Venkatesh, B., E. F. Kirkness, Y.-H. Loh, et al.; 2007; "Survey Sequencing and Comparative Analysis of the Elephant Shark (Callorhinchus milii) Genome." PLOS Biology; 5;(4); e101.

Wall, R., R. P. Ross, G. F. Fitzgerald, et al.; 2014; "Fatty acids from fish: the anti-inflammatory potential of long-chain omega-3 fatty acids." Nutrition Reviews; 68;(5); 280-289.

Wang, Q., C. N. Arighi, B. L. King, et al.; 2012; "Community annotation and bioinformatics workforce development in concert—Little Skate Genome Annotation Workshops and Jamborees." Database: The Journal of Biological Databases and Curation; 2012;(bar064.

Wang, S. P., H. Yang, J. W. Wu, et al.; 2014; "Metabolism as a tool for understanding human brain evolution: Lipid energy metabolism as an example." Journal of Human Evolution; 77;(41-49.

Watanabe, K., M. Ohno, M. Taguchi, et al.; 2016; "Identification of amino acid residues that determine the substrate specificity of mammalian membrane-bound front-end fatty acid desaturases." J Lipid Res; 57;(1); 89-99.

Watkins, P. A.; 1997; "Fatty acid activation." Progress in Lipid Research; 36;(1); 55-83.

Watkins, P. A., D. Maiguel, Z. Jia, et al.; 2007; "Evidence for 26 distinct acyl-coenzyme A synthetase genes in the human genome." Journal of lipid research; 48;(12); 2736-2750.

Weismann, A.; 1893; The germ-plasm: a theory of heredity. New York, Charles Scribner's sons.

Wolfgang, M. J., T. Kurama, Y. Dai, et al.; 2006; "The brain-specific carnitine palmitoyltransferase-1c regulates energy homeostasis." Proceedings of the National Academy of Sciences of the United States of America; 103;(19); 7282-7287.

Wu, T., L.-C. Sun, C.-H. Du, et al.; 2009; "Identification of pepsinogens and pepsins from the stomach of European eel (Anguilla anguilla)." Food Chemistry; 115;(1); 137-142.

Yakabe, E., M. Tanji, M. Ichinose, et al.; 1991; "Purification, characterization, and amino acid sequences of pepsinogens and pepsins from the esophageal mucosa of bullfrog (Rana catesbeiana)." Journal of Biological Chemistry; 266;(33); 22436-22443.

Yang, Z. and B. Rannala; 2012; "Molecular phylogenetics: principles and practice." Nat Rev Genet; 13;(5); 303-314.

Zhang, J.; 2003; "Evolution by gene duplication: an update." Trends in ecology & evolution; 18;(6); 292-298.

Zhang, J., H. F. Rosenberg and M. Nei; 1998; "Positive Darwinian selection after gene duplication in primate ribonuclease genes." Proceedings of the National Academy of Sciences; 95;(7); 3708-3713.

Zhang, Y., X. Zou, Y. Ding, et al.; 2013; "Comparative genomics and functional study of lipid metabolic genes in Caenorhabditis elegans." BMC Genomics; 14;(1); 164.

Zhao, H., S. J. Rossiter, E. C. Teeling, et al.; 2009; "The evolution of color vision in nocturnal mammals." Proceedings of the National Academy of Sciences; 106;(22); 8980-8985.

Zheng, X., I. Seiliez, N. Hastings, et al.; 2004; "Characterization and comparison of fatty acyl Δ6 desaturase cDNAs from freshwater and marine teleost fish species." Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology; 139;(2); 269-279.